Final Project Report

Grant AOARD-10-4029

Automatic Multimodal Cognitive Load Measurement (AMCLM)

June 2011

NICTA DSIM Team

Executive Summary

This report summarizes the research activities, results of the user studies and research accomplishments out of the AMCLM project in the past year. At the start of the project, we carried out a literature review on video based physiological measures of cognitive load, focusing on unobtrusive eye-activity related methods, to understand the state-of-the-art. In the mean time, we also investigated the validity of using speech formants and their fusion to measure cognitive load automatically. For the research on eye-activity based cognitive load measurement, we had examined various features, including blink latency, fixation time, saccade speed and pupil size. We further investigated the use of pupil size for automatic classification of cognitive load in different luminance conditions and under various emotional stimuli. All together, we had carried out 4 sets of user experiments to validate the research outcomes in a range of task scenarios, including Stroop test, computer-based basketball training, and mental arithmetic (summation) tasks.

In terms of concrete research outcomes, the following report and papers were published:

- Wang, Y., *Literature Review on Video Based Physiological Measures of Cognitive Workload*, NICTA Technical Report, August 2010.
- Yap, T. F., Epps, J., Ambikairajah, E. and Choi, E., “An Investigation of Formant Frequencies for Cognitive Load Classification”, *Proc. Annual Conference of the International Speech Communication Association (InterSpeech’10),* Makuhari, Japan, September 2010, pp. 2022-2025.
- Chen, S., Epps, J., Ruiz, N and Chen, F., “Eye Activity as a Measure of Human Mental Effort in HCI”, *Proc. International Conference on Intelligent User Interfaces (IUI’11),* Palo, Alto, U.S.A., February 2011, pp. 315-318.
- Xu, J., Wang, Y., Chen, F., Choi, E., Li, G., Chen, S. and Hussain, S., “Pupillary Response Based Cognitive Workload Index under Luminance and Emotional Changes”, *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI’11),* Vancouver, Canada, May 2011, pp. 1627-1632.
- Xu, J., Wang, Y., Chen, F. and Choi, E., “Pupillary Response Based Cognitive Workload Measurement under Luminance Changes”, *Proc. IFIP International Conference on Human-Computer Interaction (INTERACT’11),* Lisbon, Portugal, September 2011, to appear.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **12 AUG 2011**	2. REPORT TYPE	3. DATES COVERED

4. TITLE AND SUBTITLE **Automatic Multimodal Cognitive Load Measurement**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Fang Chen**	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National ICT Australia (NICTA),Level 5, 13 Garden Street,Eveleigh, NSW Sydney 2015,Australia,NA,NA**	8. PERFORMING ORGANIZATION REPORT NUMBER **N/A**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report summarizes the research activities, results of the user studies, and research accomplishments out of the AMCLM project in the past year. We investigated the validity of using speech formants and their fusion to measure cognitive load automatically. For the research on eye-activity based cognitive load measurement, we had examined various features, including blink latency, fixation time, saccade speed and pupil size. We further investigated the use of pupil size for automatic classification of cognitive load in different luminance conditions and under various emotional stimuli. All together, we had carried out four sets of user experiments to validate the research outcomes in a range of task scenarios, including Stroop test, computer-based basketball training, and mental arithmetic (summation) tasks.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**60**	

Contents

1 Introduction

Cognitive load (CL) refers to the amount of mental demand imposed by a particular task on a person, and has been associated with the limited capacity of working memory. However, the same task demand can affect different people in different ways, and can induce levels of perceived cognitive load that vary from one person to another. The cognitive load experienced by a person in completing a task has a major impact on her/his ability to acquire information during the task, and can severely impact the human performance if the load exceeds the mental capacity. Cognitive load measurement (CLM) therefore plays an important role in various application areas involving human-computer interface, such as air traffic control, in-car safety and electronic games. By quantifying the mental efforts of a person when performing tasks, cognitive load measurement helps predict or enhance the performance of the person and the overall system.

This project has focused on two main modalities, namely speech and eye activity, for automatic cognitive load measurement, due to the unobtrusive nature of these methods. While speech related CLM research is comparatively mature, eye activity as a physiological measure still requires much more research exploration.

2 Literature Review - Video based CLM measures

Physiological measures belong to one class of workload measurement techniques, which attempt to interpret the cognitive processes through their effect on the operator’s body state. In the past, physiological measures usually entailed invasive equipment. With the advance of sensing technologies in recent years, the measuring techniques have become less intrusive, especially those through remote sensing. As a physiological index, eye activity has been considered as an effective indicator of cognitive workload assessment, as it can be sensitive to changes of mental effort. Eye activity based physiological measures, such as fixation and saccade, eye blink, and pupillary response, can be detected unobtrusively through remote sensing.

Thus a literature review on video based physiological measures including eye activity has been carried out to understand the research landscape and the review can be found in Appendix A. This review covers imaging sensors for workload studies, pupillary, eye-blink, eye-movement and skin-temperature based workload measures. It also reviews various multimodal measures and their fusion methods. We note the increasing use of multimodal feature fusion and probabilistic modeling, and the need of user-specific modeling.

3 Speech formant based CLM

This study has allowed us to carry out a detailed analysis of the vowel-level effect of cognitive load on formants (vocal tract characteristics), together with classification results for different formant feature combinations. The corresponding paper can be found in Appendix B.

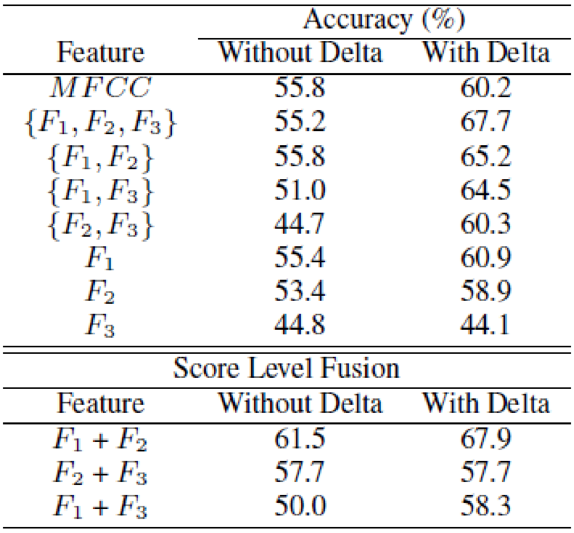
3.1 Design and procedure

The database used in this work is the Stroop test database, which consists of speech recorded from 16 randomly selected native English speakers (7 males and 9 females) while performing three tasks of varying cognitive load levels. In the low load task, speakers were asked to read aloud words corresponding to color names. In the medium load task, a mismatch was introduced between the color names and their font colors, and the speakers were asked to name the font colors instead. The high load task was similar to the medium load task except that time constraints were added to the task.

3.2 Analysis and results

A subset of the Stroop test database, comprising 4 vowel sounds (/ae/, /eh/, /iy/ and /uw/) from all speakers, was extracted. We studied how vowels change in the F1-F2 plane under different CL levels and examined the dependency of a formant frequency change on the type of vowels. Overall, we observed that F1 is increasing, while F2 is decreasing, as cognitive load is increased. Classification results performed on the Stroop test database show that formant features (F1, F2 & F3) not only have lower dimensionality, but dynamic formant features can outperform conventionally used MFCC-based features by a relative improvement of 12%. The classification results are shown in Table 1.

Table 1. 3-class cognitive load classification using formant and MFCC features.



	Accuracy (%)	
Feature	Without Delta	With Delta
$MFCC$	55.8	60.2
$\{F_1, F_2, F_3\}$	55.2	67.7
$\{F_1, F_2\}$	55.8	65.2
$\{F_1, F_3\}$	51.0	64.5
$\{F_2, F_3\}$	44.7	60.3
F_1	55.4	60.9
F_2	53.4	58.9
F_3	44.8	44.1

	Score Level Fusion	
Feature	Without Delta	With Delta
$F_1 + F_2$	61.5	67.9
$F_2 + F_3$	57.7	57.7
$F_1 + F_3$	50.0	58.3

4 Eye activity based CLM

This study researches into 8 eye activity based features, spanning eye blink, pupillary response and eye movement information. Correlation analysis between various pairs of features suggests that significant improvements in discriminating different effort levels can be made by combining multiple features. A conference paper has been published for this study and it can be found in Appendix C.

4.1 Design and procedure

A computer-based training application, running on a tablet monitor, was designed for basketball players to learn playing strategies by observing team player positions in basketball game videos. The goal of this task was to detect and identify defenders and attackers during a video clip of an actual game, and recall their positions around the ball at the end of each 15-second clip.

Subjects were instructed to watch a game video clip and recall player positions by writing them down on a blank on-screen basketball court schematic using simple signs: crosses and circles. They completed 6 sub-tasks for each low, medium and high level of mental demand, with a few minutes break between each level. All participants completed 8 sessions in different days, and here we consider one of those 7 sessions for data analysis.

Task difficulty levels were varied by the number of player positions to be recalled. In the low cognitive load level, 3 player positions were required, while 6 positions were required by the medium level and all 10 positions in the high level.

Twelve paid male recreational basketball players, each with more than two years' experience, aged 19-36, completed this experiment. Eye activity was monitored using an ASL Eye-Trac 6 head mounted eye tracker system. Subjects were free to move their head but instructed to keep their eyes within the screen display range.

4.2 Analysis and results

Eight dependent variables were employed to measure the mental effort: blink latency (BL), blink rate (BR), mean pupil size (MPS) in the time between 2s preceding and after the game video ended, standard deviation of pupil size (SPS) in the 4-second period, fixation time (FT) , fixation rate (FR), saccade size (SSI) and saccade speed (SSP). Table 2 shows the results of paired t-test conducted on these measures.

Table 2. Paired t-test for eye activity based measures

	BL	BR	MPS	SPS	FT	FR	SSI	SSP
t(5)	1.94	3.64	4.22	0.62	2.41	2.95	4.68	3.56
P_{value}	0.109	0.014	0.008	0.557	0.060	0.031	0.005	0.016

In regards to blink activity, both blink latency and blink rate display clear mental effort related variations. Pupil size was measured from 2 seconds before and 2 seconds after the clip, involving mostly recall in this period, during which sustained working memory is heavily involved. The average pupil size for the two difficulty levels shows a significant effect as opposed to the standard deviation of pupil size, which indicates that in some cases pupil size is larger in a more difficult task level but shows less fluctuation. Meanwhile, fixation duration and fixation rate results indicate that significantly more attention is needed when the task is more complex. In addition, saccade speed and especially saccade size appear to have been highly discriminatory parameters.

5 Pupillary response based CLM

Two studies were carried out to investigate how the relationship between pupil size and cognitive load may be affected by various factors unrelated to workload, including luminance condition and emotional arousal. The corresponding papers can be found in Appendices D and E respectively.

5.1 Workload measurement under luminance changes

5.1.1 Design and procedure

Each subject is requested to perform arithmetic tasks under different luminance conditions. The arithmetic tasks have 4 levels of difficulty, and each level of task difficulty is combined with 4 levels of background brightness, which results in 16 different trial types in total.

For each arithmetic task, each subject is asked to sum up 4 different numbers sequentially displayed on the center of the screen, and then choose the correct answer on the screen through mouse input. The task difficulty depends on the range of numbers. For the first (lowest) difficulty level, each number is binary (0 or 1); for the second difficulty level, each number has 1 digit (1 to 9); for the third difficulty level, each number has 2 digits (10 to 99); for the fourth (highest) difficulty level, each number has 3 digits (100 to 999). Each number will be displayed for 3 seconds, and there is no time constraint for choosing the answer. Before the first number appears, different number of “X” will be displayed at the center of the screen for 3 seconds. The number of “X” corresponds to the number of digits for each arithmetic task.

During the experiment, the luminance condition is varied when each subject performs arithmetic tasks. To produce different levels of luminance condition, luminance (grayscale value) of the background are set as 32, 96, 160 and 224 for the four levels of background brightness (L1, L2, L3, and L4), respectively. Black background will be displayed for 6 seconds before each arithmetic task.

The experiment starts with a practice trial of which the data is not analyzed. Subsequently a one-minute resting data with black background is recorded before the test trials start. There are two tasks for each trial type, which results in 32 arithmetic tasks for each subject in the experiment. The tasks are presented randomly during the experiment. Once the subject finishes all the tasks, another one-minute resting data is also recorded. The whole experiment lasts about 25 minutes for each subject.

Thirteen 24-to-46-year-old male subjects have been invited to participate in the experiment. All the subjects have normal or corrected-to-normal vision. Each subject receives a small-value reward for his participation.

5.1.2 Analysis and results

For each subject, the pupillary response data of every arithmetic task during the experiment is examined. For a coarse-grained analysis, the average pupil diameter from the whole task period is used to characterize the cognitive workload. Together, background brightness and cognitive workload could affect the pupil diameter. It can be observed that the pupil diameter at the highest task difficulty with highest background brightness is, in fact, smaller than that at the lowest task difficulty with lowest background brightness. This observation is consistent with previous empirical study that luminance conditions take priority over cognitive demands in pupil diameter changes. Thus it is difficult to directly use the average pupil size or dilation to measure cognitive workload in the experiment.

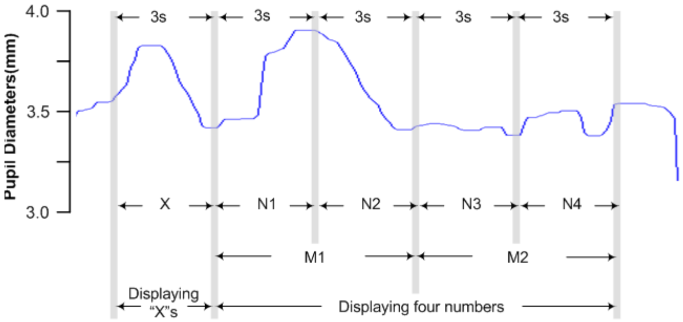


Figure 1. The setting of task intervals for fine-grained analysis

To overcome this problem, we propose a fine-grained analysis of pupillary response by dividing the task period into smaller-size intervals (see Figure 1). It is expected that the dynamic characteristics of cognitive process can be reflected by the fine-grained measures of pupillary response, which will improve cognitive workload measurement under complex environments. The corresponding measurement values significantly correlates to the task difficulty level (F=3.93, $p<0.05$ in ANOVA test).

5.2 Workload measurement under luminance and emotional changes

5.2.1 Design and procedure

The subjects perform arithmetic tasks under changes of luminance condition and emotional arousal simultaneously. The whole experiment consists of three parts and lasts about 15 minutes. In the first part, the subject is asked to sum up numbers with blank background (black screen). In the second and third parts, the subject is asked to sum up numbers with pleasant and unpleasant background images shown on the screen. Different task difficulty levels and background conditions are employed to manipulate the cognitive workload, as well as background luminance and emotional arousal during the experiment. The setting of arithmetic task and its difficulty level are the same as those described in the previous section.

To vary both the luminance condition and emotional arousal, pleasant and unpleasant background images are shown on the screen when the subject performs arithmetic tasks in the second and third parts of the experiment. A background image will be displayed for 6 seconds before each arithmetic task. Subsequently, the subject will perform the arithmetic task with the background image remaining on the screen. Eight pleasant images (mean valence/arousal = 7.1, 5.7) and eight unpleasant images (mean valence/arousal = 2.8, 4.8) are selected from the IAPS database. The mean luminance of the images ranges from 53 to 174.

One minute resting data with black screen is recorded at the beginning and the end of the whole experiment for each subject. There are 8 arithmetic tasks randomly given in each experiment part (2 for each difficulty level). Twelve 24-to-35-year-old male subjects have been invited to participate in the experiment.

5.2.2 Ongoing analysis

To investigate the feasibility of robustly measuring cognitive workload even under the effects of noisy factors including luminance changes and emotional arousal, a simple difference feature (the difference of the average pupil diameter between the first half and second half interval of the task) is employed to characterize the cognitive workload changes. The distribution of normalized feature values for different difficulty levels from all the pupillary response data is depicted in Figure 2. The feature value increases as the task difficulty level (F>8, p<0.01 in ANOVA test).

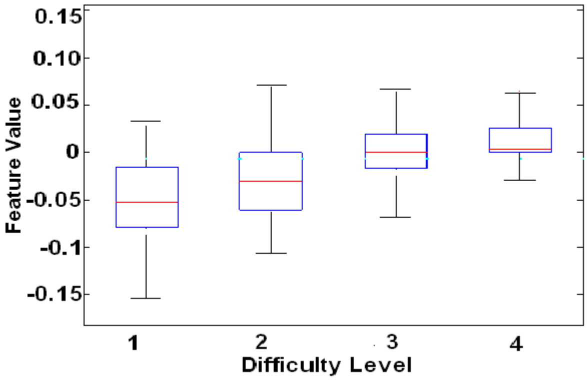


Figure 2. Box plot of feature values (sample minimum, lower quartile, median, upper quartile, and maximum) corresponding to different task difficulty levels

Our current work focuses on developing machine learning algorithms that can automatically find optimal features for robust workload measurement under noisy factors. There are quite a few systematic ways for solving this optimization problem, and Boosting is one popular algorithm that is suitable in this instance. Boosting is a type of learning algorithm, which creates a classifier that can predict the labels of unseen data based on the given examples and their labels. In its original form, the Boosting algorithm is used to form a strong classifier from a set of weak classifiers. A strong classifier is defined as a classifier that correlates arbitrarily well with the true classification, whereas a weak classifier only correlates slightly with the true classification. However it can also be used as a feature selection scheme if we relate each weak classifier to a single feature. For example, we can define a weak classifier $h_j(x)$ that consists of a feature f_j, a threshold θ_j, and the parity $p_j = \pm 1$, which indicates the following simple classification rule: if $p_j f_j(x) \le p_j \theta_j$ then $h_j(x) = 1$, otherwise $h_j(x) = 0$. The Boosting algorithm then creates a strong classifier $H(x) = \sum_j \alpha_j h_j(x)$ by selecting $h_j(x)$ iteratively from a pool of weak classifiers, and each $h_j(x)$ is weighted by α_j, which relates $h_j(x)$'s classification accuracy on the examples. Additionally, the examples are reweighted so that future weak classifiers can focus on the examples misclassified by the previous classifiers. In this work, each extracted feature can be viewed as a weak classifier consisting of a time interval vector $\vec{T_j}$, a threshold and also a parity value. Currently $\vec{T_j}$ is set heuristically using the first and second half of each task. To improve accuracy of cognitive workload indexing, the optimal set of weak classifiers (features) can be obtained through the Boosting algorithm.

6 Conclusion

Formant frequencies have been studied, firstly, trying to understand the effect of cognitive load on formants, and hence the speech production system, and secondly, finding effective features for cognitive load classification. In general, F1 is found to be increasing, and F2 decreasing, when cognitive load is increased. Additionally, we have also found that formant frequencies exhibit vowel-specific shifts in their mean values under different cognitive load conditions. For classification purposes, not only formant features have lower dimensionality compared with the baseline system, they also outperform the baseline by a relative improvement of 12% when dynamic information is incorporated. Future work will include an investigation of different features to capture dynamic formant information, and their fusion with other vocal source features for improved cognitive load classification.

Eye activity features have been shown to each provide significant discriminative power between different levels of induced cognitive load. Combination of these features has a distinct advantage as an objective measure of human cognitive load, as different inhibitory mechanisms require mental effort for eye functions that, when combined, provide rich and possibly complementary information about cognitive load. In turn, we may able to improve our understanding of human cognitive load in real time, which may prove significant in the design and evaluation of usable, intelligent adaptive interfaces. The experimental results also demonstrate the feasibility of cognitive workload measurement under complex environments using the fine-grained analysis. Our future work will be applying machine learning techniques to improve fine-grained analysis for cognitive load measurement. Going forward, we will investigate the robust fusion of different modalities that incorporates some other information streams, such as skin conductance and electroencephalogram (EEG).

Appendix A

Literature Review on Video Based Physiological Measures of Cognitive Workload

Yang Wang
August 2010
Making Sense of Data
NICTA

Contents

1 Introduction

Cognitive (mental) workload is an important issue in various application areas such as human computer interaction, adaptive automation and training, traffic control, performance prediction, and driving safety (Byrne and Parasuraman, 1996; Coyne et al., 2009; Grootjen et al., 2007; Wilson and Russel, 2006). Although numerous approaches have been developed to study cognitive workload or understand how hard the brain is working under various situations, it is still difficult to examine the cognitive workload of a person: “workload is a multidimensional, multifaceted concept that is difficult to define. It is generally agreed that attempts to measure workload relying on a single representative measure are unlikely to be of use” (Gopher and Donchin, 1986). Both theories and models have been proposed to explain cognitive workload. The multiple resource theory models cognitive resource of a person with three different dimensions: perceptual modality, information code, and processing stage (Wickens, 2002). On the other hand, the cognitive load theory models the interaction between limited working memory and relatively unlimited long term memory during the learning process (Sweller, 1988). The theory distinguishes between three types of cognitive workload: intrinsic load, extraneous load, and germane load. The first type is associated with the nature of learning material, while the latter two are influenced by instructional design (Paas et al., 2003).

When a subject or operator is required to perform a given task, cognitive workload could be viewed as the interaction between the demands of the task and the capacity of the subject (Cain, 2007). Such point of view highlights two key issues of mental workload, the subject’s capacity and the task demands. The mental workload of a subject tends to increase when the cognitive capacity becomes low, and it tends to increase when the task demands become high. It should be noticed that both subject’s capacity and task demands are not necessarily constant values and they may change over time. The capacity of an operator may increase or decrease due to various factors such as training, fatigue, and environment. During a task, an operator can also experience varying levels of workload according to the task difficulty at different stages.

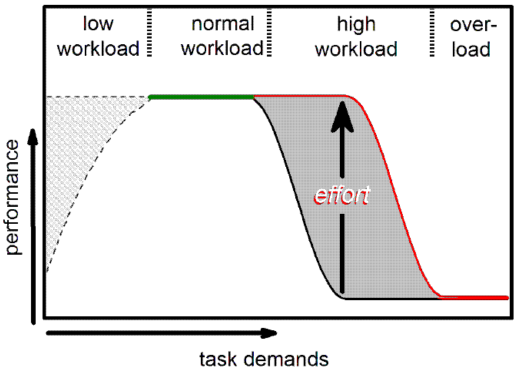


Figure 1. The relationship between task demands, performance, and workload (Veltman and Jansen, 2006).

In recent decades, a great variety of measuring techniques, from simple ones such as questionnaires to complex ones such as functional braining imaging, have been developed to study cognitive workload (Gingell, 2003; Just et al., 2003; Wierwille and Eggemeier, 1993). Generally, these measuring techniques can be divided into three categories: subjective rating, performance measure, and physiological measure (Hart and Staveland, 1988; O'Donnell and Eggemeier, 1986; Wilson et al., 2004). Comparing with subjective rating, the latter two categories provide approaches to assess mental workload in an objective way. One main advantage of objective measurement is that it will not disturb the operation of the subject during the task execution. For performance based measure, the relationship between workload and performance is shown in Figure 1. An operator's performance could be maximized if the task just requires normal mental workload. Meanwhile the performance tends to decline when the task demands become high or even exceed the capacity of the operator. The performance is also influenced by various factors such as attention, expertise, experience, stress, and motivation.

With the advance of modern sensor technologies, more and more physiological measures have been developed for the assessment of cognitive workload. Popular physiological measures used in workload studies include brain wave, eye activity, respiration, heart rate, and speech, etc (Fournier et al., 1999; Scerbo et al., 2001; Yin et al., 2008). Among these techniques, video based workload measures, especially the ones through remote sensing, have attracted increasing attention since they can provide physiological evaluation of cognitive state in a non-intrusive and non-obtrusive way.

Although various studies exhibit the effects of mental workload on physiological measures, no single physiological measure will be sufficient to comprehensively characterize the workload, especially in the case of multidimensional task and/or dynamic circumstances. On the other hand, changes in physiological measures may take place due to a lot of other aspects, such as engagement, fatigue, stress, and environment. Mental workload is just one of these factors influencing physiological measures.

2 Video Based Workload Measures

For the convenience of cognitive workload measurement in different experiments, sensors are selected by the following three usability criteria (Voskamp and Urban, 2009): non-intrusiveness, non-obtrusiveness, and simplicity. Usually a subject does not prefer a device that may invade the human body in any way. Ideally, the applied sensor will not interrupt the operator during the task execution. Moreover, it should not require much effort or training to gather the measurement data.

For the effectiveness of mental workload measures in various cognitive tasks, sensors should also meet the following three technology criteria: sensitivity, efficiency, and compatibility. The selected sensor is required to provide data that is highly correlated to cognitive workload. For online or interactive systems, the collected data needs to be transferred and processed in real-time. When multiple sensors are applied, the sensors should be easily combined with each other.

2.1 Imaging sensors for workload studies

Based on the sensor selection criteria for cognitive workload study, video camera or imaging sensors have attracted increasing interests during the development of workload measurement techniques. One valuable type of physiological measure involves workload effects on activities of human eye (Sirevaag and Stern, 2000). Especially, video based eye tracker becomes a popular approach for cognitive workload evaluation due to its sensitivity and convenience. Eye tracking data provides important information about human brain activity and autonomic nervous system, and it is highly correlated with subject's mental workload. The visual information is acquired in a non-intrusive (particularly with remote systems, see Figure 2) and continuous way without interfering user's activity during the task performance. Moreover, the video sequences of eye tracking data can be captured with high frame rate (more than 30 frames per second) and processed in real-time.

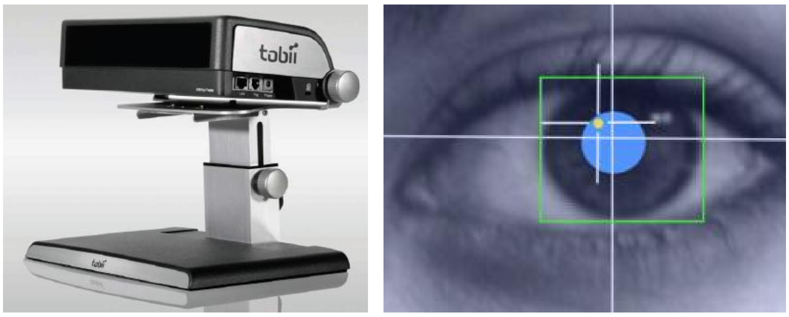


Figure 2. Eye tracker and eye tracking data.

Another type of imaging data, facial skin temperature, has been utilized as a physiological measure in mental workload studies as well. Facial skin temperature shows significant correlation to changes of mental status (Veltman and Vos, 2005). During the task execution, the autonomic nervous system of the subject causes the redistribution of blood flow. Consequently, it will result in the change of local skin temperature. With the use of thermal infrared camera (see Figure 3), the remote sensing of skin temperature can be achieved through measuring the infrared emitted from human body.

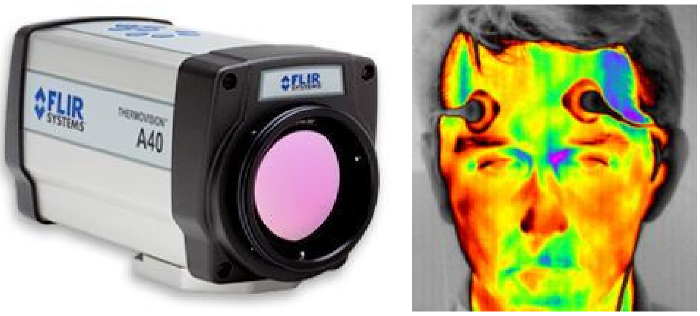


Figure 3. Infrared camera and thermal imaging data.

On the other hand, advanced brain imaging techniques, such as magnetic resonance imaging (MRI) and near-infrared (NIR) neuroimaging, have also been employed to detect changes in cognitive workload (Callicott et al., 1999; He et al., 2007; Izzetoglu et al., 2005). However, due to the constraint of sensing technology and device, in practice it is hard for those sensors to capture the imaging data in a convenient and non-obtrusive way, which limits their usability as physiological measures of mental workload.

2.2 Video based measures in cognitive tasks

Eye tracking data provides rich information for cognitive workload assessment. Physiological workload features related to eye activity can be categorized into three classes: eye blink based measures, eye movement (saccade and fixation) based measures, and pupillary response based measures.

In addition to eye activity and facial skin temperature, physical features of human behaviour such as head movement, hand gesture, and facial expression, which can be detected in a non-contact way using video camera, also provide useful information about changes in mental states (Grootjen et al., 2006). Since such physical behaviour measures are relatively less sensitive to cognitive workload, they are usually integrated with other physiological measures to achieve satisfactory performance. Table 1 lists popular video based measures that have been used in cognitive workload studies. Workload research groups working on video based physiological measures include Air Force Research Laboratory, Naval Health Research Center, Human Factors Group of Federal Aviation Administration, TNO Human Factors Research Institute, etc.

To study the effects of mental workload on physiological measures, various cognitive tasks have been designed and tested in both laboratory and real world. The performed tasks include visual, auditory, arithmetic, executive, and complex ones such as driving, traffic control, and flight. Sometimes dual tasks are performed in the workload experiments. It should be noted that multitasking is common for human activities under laboratory and real world environment. For example, even a simple auditory addition task will involve both verbal processing and arithmetic processing; a driving task can be decomposed into at least two subtasks (visual and memory) that require the driver to keep the vehicle on the road and remember the route to the destination. When multiple physiological measures are available, it will be sensible to consider the embedded multimodal information with a composite index for mental workload evaluation (Sciarini and Nicholson, 2009).

Table 1. Video based workload measures

Category	Measure	Explanation
Eye blink based	Blink frequency	The rate of blink times over a time period
	Blink duration	The time interval during the closure of eye
	Blink interval	The time interval between two successive eye blinks
Eye movement based	Fixation number	The times of fixation
	Fixation frequency	The rate of fixation times over a time period
	Fixation duration	The time interval during a fixation
	Saccade rate	The rate of saccade times over a time period
	Saccade extent	The angular distance within a saccade
	Saccade duration	The time interval during a saccade
	Saccade velocity	The angular velocity within a saccade
	Scan path	The trajectory of eye gaze
	Vergence angle	The gaze difference between left eye and right eye
Pupillary response based	Pupil dilation	The increase of pupil size comparing with baseline
	Percentage change in pupil size	The rate of pupil dilation over baseline pupil size
	Index of cognitive activity	Based on changes in pupil dilation (Marshall, 2002)
	Power spectrum	The power spectrum of pupil size data
Skin temperature based	Nose temperature	
	Forehead temperature	
Physical behaviour based	Head movement	
	Facial expression	
	Hand movement	

3 Pupillary response based measures

The correlation between pupillary response and changes in mental workload has been observed for decades (Beatty, 1982). It is known that human eye is regulated by the autonomic nervous system, and pupil diameter will decrease or increase based on autonomic response. Increased pupil diameter is usually observed with an increase in workload demand. Generally, pupil dilation is an important physiological measure of mental efforts and has been widely applied as an effective indicator of cognitive workload.

3.1 Correlation to workload in visual task

Backs and Walrath (1992) evaluated the changes in mental workload when utilizing colour coding for symbolic tactical display in a visual search task. Participants were required to abstract different types of information from the display with varying symbol density. Two pupillary response measures, pupil dilation and constriction-dilation difference, were collected as physiological indices of visual workload. It was found that pupillary response was not only affected by display parameters such as colour coding and symbol density, but also sensitive to the information processing demands of the visual task.

In the experiment of a visuospatial task with varying target density (Van Orden et al., 2001), changes in eye activity based physiological measures were examined during the task. Pupil diameter, together with blink frequency and fixation frequency, were found to be the most relevant eye activity features regarding the target density. Moreover, in the experiment of cognitive task and visual search task (Recarte et al., 2008), the analysis results exhibited that pupil dilation could effectively measure the mental efforts during the cognitive tasks, and it could be used as a physiological predictor of visual impairment as well.

Verney et al. (2001) investigated task-evoked pupillary response in the experiment of a visual backward masking task. The experimental results showed that pupil dilation response became significantly greater during the task condition than during the passive condition of stimulus viewing. Comparing with the non-mask condition, the pupil dilation exhibited significantly increase under the masking condition, especially when the interval between target and mask stimuli was prolonged. As pupillary dilation increased when resource allocation became intensive in the visual task, the experiment demonstrated that the mask could demand extra processing resources when it followed the target by prolonged interval.

Both time domain and frequency domain of physiological data provide useful information for mental workload estimation. The power spectrum of pupillography, especially the band of lower frequency, could be employed as a physiological measure of mental activity as well. Nakayama and Shimizu (2004) studied the frequency information from the task-evoked pupillary response. In the experiment, participants performed visual following task together with/without oral calculation task under different difficulty levels. Pupil size was recorded as physiological measurement during the task performance. It was found that for the oral calculation task, the power spectrum density of pupil size data increased with higher task difficulty level in the band of 0.1-0.5 Hz and 1.6-3.5Hz, which was consistent with the changes in average pupil size.

Pupil dilation is known to exhibit effects of both the illumination condition of the visual field and the cognitive workload of the person while performing a visual task. Pomplun and Sunkara (2003) investigated effects of cognitive workload and display brightness on pupil dilation and their interaction in the experiment of a gaze-controlled human-computer interaction task. During the visual task, three levels of task difficulty were combined with two levels of background brightness (black and white). The experimental results showed that under both black and white background conditions, the pupil area exhibited significant increase when workload demands became higher. However, under bright background even the pupil area corresponding to high level of task difficulty was significantly smaller than the pupil area corresponding to low level of difficulty under black background. Hence comparing with the task difficulty, the background brightness actually resulted in greater variation of pupil area.

3.2 Correlation to workload in driving task

Marshall (2002) proposed a physiological measure of cognitive workload, index of cognitive activity (ICA), from changes in pupil dilation. ICA would measure abrupt discontinuities in pupil diameter and try to separate pupil's reflex reaction to changes in light from the reflex reaction to changes in workload. In the cognitive workload study with a simulated driving task (Schwalm et al., 2008), the experimental results showed that ICA increased when workload demands became high, which was induced by performing lane change manoeuvre or additional secondary task. The study exhibited the feasibility of ICA as a physiological measure of mental workload while driving.

In a dual task experiment, Tsai et al. (2007) examined pupillary response when subjects performed driving task and auditory addition task simultaneously. It was found that pupil dilation was significantly greater when subjects were performing well in the auditory task than when subjects were performing poorly.

In another experiment of dual task, Palinko et al. (2010) also studied the pupillary response with remote eye tracker. The subjects performed simulated vehicle driving as well as spoken dialogues. In the experiment, pupil size data acquired from remote eye tracker was used for the evaluation of the driver's cognitive load. During the task, the physiological measure based on pupillary response exhibited significant correlation to those measures based on driving performance. A pupillary response based measure of cognitive load, mean pupil diameter change rate, was proposed to analyse workload changes with small time scales. The experimental results demonstrated the reliability of physiological measures obtained through remote eye tracking for cognitive load estimation.

3.3 Correlation to workload in arithmetic/memory task

Murata and Iwase (1998) assessed mental workload based on the fluctuation of pupil area. In the experiment, a mental division task and a Sternberg memory search task were carried out with the controlling of respiration. During the task, the number of digits and the size of memory set were used to manipulate the mental workload level induced by task demands. For each subject, the autoregressive power spectrum of pupil area was used for cognitive workload assessment. It was found that the ratio of power at low frequency band (0.05-0.15Hz) over power at high frequency band (0.35-0.4Hz) increased with higher level of task difficulty for both the arithmetic task and the memory task. The experimental results indicated that the fluctuation rhythm of the pupil area could be used as an effective physiological index to evaluate mental workload.

Klingner et al. (2008) examined the pupil measuring capability of video based eye tracker for cognitive workload evaluation. In the experiment of several tasks including arithmetic and memory ones, subtle changes of pupil size in the task-evoked pupillary response were detected using remote eye tracker. Comparing with the results in earlier studies, it was found that cognitive workload could be effectively measured through remote eye tracking. Moreover, the experimental results exhibited the feasibility of analysing the timing and magnitude of short-term pupillary response based on the collected eye tracking data, which could provide more details about changes in cognitive workload.

3.4 Correlation to workload in other tasks

In an early study, Beatty (1982) investigated task-evoked pupillary response in the experiments of various tasks such as language processing, reasoning, and perception. Pupil dilation was exhibited as a reliable physiological measure of mental state or processing load during the task performance. Similarly, in the recent experiment of a combat management task involving target identification (Greef et al., 2009), pupil dilation also increased when cognitive workload became high.

In the experiment of air traffic controller task (Ahlstrom and Friedman-Berg, 2006), mean pupil diameter was employed as the physiological measure of mental workload. It was found that comparing to when using a dynamic forecast tool, the mean pupil diameter became significantly larger when using a static forecast tool. The experimental results indicated that the use of static tool led to higher cognitive workload. In another experiment of a video game task (Lin and Imamiya, 2006), it was also found that pupil size increased when task difficulty changes from low level to high level.

For interruption management in interactive systems, notifications delivered during the period of lower mental workload would become less interruptive (Iqbal et al., 2004). Bailey and Iqbal (2008) empirically examined changes in mental workload during goal-directed interactive tasks including reading comprehension, mathematical reasoning, product searching, and object manipulation. Percentage change in pupil size was used as the task-evoked pupillary response for continuous workload measurement. The experimental results showed that workload would decrease at subtask boundaries, and the decrement would be greater at boundaries when the operators accomplished large chunks of the interactive task. For operators of interactive systems, pupillary response was exhibited to be a meaningful index of mental workload during the execution of a hierarchical task.

Although mental workload has been exhibited to decrease at subtask boundaries, it has not been examined for subtasks requiring different devices such as notebook computer and mobile phone. Tungare and Perez-Quinones (2009) proposed to study the changes in mental workload for multi-device personal information management. In an ongoing experiment, participants would perform information collection tasks using different devices. Pupil diameter would be monitored to provide continuous measurement of workload.

Existing software analysis tools usually can generate the graph of pupillary response over time and playback the video of user's screen interaction, but may not allow the response data to be interactively explored with regard to the task execution model. To facilitate analysis of pupillary response data in relation to the hierarchical structure of the task, Bailey et al. (2007) developed an interactive analysis tool to analyse mental workload if the task could be decomposed into hierarchical subtasks. The workload data was precisely aligned to the corresponding task execution model during the analysis.

4 Eye blink based measures

Pervious research work has exhibited that eye blink is a useful measure of mental workload (Fogarty and Stern, 1989), especially for workload demands associated with visual tasks. In several experiments using either electrocculogram (EOG) or video eye tracker, blink rate decreases with an increase in cognitive workload; increase of blink interval is observed with increased mental workload; meanwhile blink duration tends to decrease against more intense processing load. Such blink based physiological response help human eye to save more time to handle visual information during the task performance.

4.1 Correlation to workload in visual task

Van Orden et al. (2001) investigated changes in various eye activity based measures in a visuospatial memory task with varying target density. Two eye blink based measures, blink frequency and blink duration were monitored during the task. In the experiment, subjects were required to recognize and remember each target's identification (friend or enemy) on the display for appropriate action (fire or not) when the targets were approaching. It was demonstrated that both blink frequency and blink duration declined with increasing target density during the visuospatial memory task.

Recarte et al. (2008) examined the concurrent validity of eye activity based physiological measures for mental workload evaluation. The participants performed single cognitive task and dual task (cognitive task and visual search) in the experiment. Under single task condition, blink rate and pupil dilation showed concurrent validity for mental workload assessment. However, the blink rate exhibited opposite effects under the dual task condition. The blink rate increased when the mental workload of cognitive task became high, meanwhile the blink rate deceased when visual demands became high.

Startle eye blink reflex is also affected by workload demands during visual task. Neumann (2002) studied changes in startle blink during a continuous visual task with different levels of mental workload. In the experiment, subjects performed a single task of visual horizontal tracking or a dual task of both visual horizontal tracking and visual gauge monitoring. The startle blink reflex was evoked by a noise burst during the task execution. Experimental results exhibited that compared with pre-task and post-task conditions, startle blink was suppressed during the task performance. Moreover, compared with the single task condition, the suppression became more significant under the dual task condition. The startle blink rate and other measures such as subjective rating and heart period showed concurrent validity for different workload levels, which indicated that startle blink could be a useful physiological measure of mental workload during the visual task.

4.2 Correlation to workload in flight task

Veltman and Gaillard (1998) investigated the sensitivity of various physiological indices, including eye blinks, in simulated flight tasks. In the experiment, subjects simultaneously performed a continuous memory task during the flight. Eye blink based measures including blink interval, blink duration, closing time and amplitude were monitored during the experiment. Comparing with the measurement data during rest status, blink interval increased and blink duration decreased when subjects performed flight tasks. In addition, blink interval increased and blink duration decreased when subjects were processing more visual information during the flight. On the other hand, the experimental results also showed that the blink interval decreased with increasing difficulty level of the memory task. The decrement was probably due to sub-vocal activity that stimulated the muscles of eyelid and resulted in increased eye blinks.

Similar results were found by Wilson (2002) in the experiment of real flight task. For each pilot, eye blink was recorded as one physiological measure during a flight with both visual rule and instrument rule conditions. The results showed that blink rate decreased when the segments of flight became more visually demanding. In the experiment, each pilot repeated the same task to examine the reliability of the physiological measures, and similar response data was obtained for the two rounds.

4.3 Correlation to workload in traffic control task

Brookings et al. (1996) examined the sensitivity of physiological response to changes in cognitive workload during simulated air traffic control task. In the experiment, eye blink rate exhibited significant effects of task difficulty. The level of task difficulty was manipulated by varying traffic volume and traffic complexity. Eye activity based physiological measures including blink rate were monitored during the traffic control task. The experimental results showed that blink rate decreased with increasing cognitive load.

Ahlstrom and Friedman-Berg (2006) investigated the effect on cognitive workload with/without the use of weather display during air traffic controller task. In the experiment, blink frequency and blink duration were used as two of the physiological workload measures. It was found blink duration became significantly shorter when controllers operated without using weather display, corresponding to a higher level of controller workload. The experimental results also indicated that comparing with subject rating, eye activity based features was relatively sensitive to the variation of mental workload at system interaction stages.

4.4 Correlation to workload in other tasks

In an experiment of dual task, Tsai et al. (2007) investigated changes in eye activities while subjects performed driving task and paced auditory serial addition task. In the experiment, two eye blink based physiological measures, blink frequency and blink duration were recorded. Experimental results exhibited that comparing with the measurement data in the single task of driving, blink frequency increased in the dual task of both driving and auditory addition. In another experiment of complex decision making task (Boehm-Davis et al., 2006), the results exhibited that eye blinks would be suppressed during cognitive processing comparing to when the processing was accomplished.

Ryu and Myung (2005) employed multiple physiological measures to evaluate the mental workload in a dual task with different difficulty levels. In the experiment, the subjects simultaneously performed a tracking task of simulated instrument landing and mental arithmetic task of adding pairs of numbers. Eye blink interval was employed as one physiological measure for mental effort assessment in both tasks. It was found that the blink interval revealed sensitivity to the changes in mental workload for the tracking task, but not for the arithmetic task.

5 Eye movement based measures

Eye movement mainly consists of two forms of activity: fixation and saccade. During the visual scan, human eyes are directed to interesting areas where fixations occur. A fixation is a steady focus of the eye, inputting detailed information of the visual stimulus into human vision system. The movement from one fixation stimulus to another is defined as a saccade. Previous studies revealed correlations between changes in mental workload and properties of eye movement (May et al., 1990). For example, the increase in fixation time has been observed with the increase of mental workload. In several experiments saccade based measures such as saccade speed also exhibited sensitivity to changes in mental workload.

5.1 Correlation to workload in visual task

In the task of visual search of symbolic displays (Backs and Walrath, 1992), number of eye fixation, fixation duration, and fixation frequency were employed as eye movement based physiological indices. It was found that the number of eye fixations was affected by both colour coding and symbol density. In the experiment participants made fewer fixations to search colour-coded displays than monochrome displays, and fewer fixations to search low-density displays than high-density displays. Moreover, compared to when searching monochrome displays, fixation duration became shorter and fixation frequency became higher when searching colour-coded displays.

In the visuospatial memory task of target identification (Van Orden et al., 2001), the task difficulty was manipulated by varying the number of targets presented on the display. Physiological measures including fixation frequency, dwell time, and saccade extent were recorded for each participant in the experiment. It was found through nonlinear regression analysis that among the eye movement based measures, fixation frequency revealed significant correlativity to the target density in the visuospatial task.

Frequency information of eye movement also provides a useful physiological index of mental workload. Nakayama and Schimizu (2004) performed frequency analysis of eye movement data in both single task of ocular following and dual task of ocular following and oral calculation. After correcting the artefacts of eye blinks in saccadic eye movement, cross spectrum density, which exhibits relationship between horizontal and vertical eye movement, was employed as a workload measure. Given the eye movement data of different task difficulty levels, the cross spectrum density exhibited significant differences between them in the frequency band of 0.6-1.5Hz.

5.2 Correlation to workload in driving/riding task

In the experiment with a dual task of driving and auditory addition (Tsai et al., 2007), three physiological measures of eye movement, including fixation frequency, fixation duration, and horizontal vergence, were assessed as the indicator of cognitive workload. Comparing to when the subjects performed poorly in the auditory task, the horizontal vergence increased when subjects performed well. Although there was no significant change in fixation frequency, it was found that fixation duration before incorrect responses of auditory addition were significantly shorter than fixation duration before correct responses in the dual task. The experimental results indicated that eye movement based measures could be utilized to both evaluate cognitive load and predict task performance in real-time.

In the experiment of motorbike riding task, Di Stasi et al. (2009) studied the relationship between cognitive workload and risk behaviour. Eye movement based measures including saccadic number, saccadic amplitude, saccadic duration, peak saccadic velocity, fixation number, fixation duration were used as physiological indices of mental workload. The experimental results showed that comparing with low-risky participants, the cognitive workload became higher for high-risky participants, meanwhile the peak saccadic velocity could be used as a reliable physiological index of risk behaviour.

5.3 Correlation to workload in traffic control task

In an experiment of air traffic control task (Brookings et al., 1996), subjects performed simulated traffic control tasks with varying traffic volume and traffic complexity. Two eye movement based workload measures, saccade rate and amplitude, were recorded together with other physiological measures during the control task. However, the saccade measures did not demonstrate significant effects of task difficulty or traffic complexity in the experiment.

Di Stasi et al. (2010) studied the effects of mental workload on eye movement based indices in simulated air traffic control task. In the experiment, participants performed multitasks with three levels of task difficulty according to the cognitive resource requirement. Three eye movement based physiological measures, saccadic amplitude, saccadic duration, and saccadic peak velocity, were recorded using video eye tracker. Experimental results showed that the peak velocity decreased with increasing task difficulty, indicating the sensitivity of saccadic movement to changes in mental workload.

5.4 Correlation to workload in other tasks

Lin and Imamiya (2006) explored the multimodal information of workload measures for usability evaluation. Multiple physiological measures, including fixation number, fixation duration, scan path length, are recorded to estimate cognitive workload when subjects were performing a video based action-puzzle game task. In the experiment, eye movement data exhibited correlation to mental workload level. It was found that mean values of three eye movement based workload measures increased when the task difficulty changed from low level to high level. Saccade speed also exhibited correlation with heart rate variability during the game task. Moreover, a composite physiological measure combining eye fixations with hand movement (mouse clicks) was proposed to improve the evaluation of task performance.

In the experiment of a combat management task requiring target identification and weapon engagement, Greef et al. (2009) investigated three aspects of eye movement, fixation time, saccade distance, and saccade speed, for objective assessment of mental workload. To examine their correlativity with changes in workload, these features of eye activity were monitored by video eye tracker under different levels of mental workload. The experiment results exhibited that fixation time significantly increased when the mental workload became high. Meanwhile saccade distance and saccade speed did not exhibit any significant effects.

6 Skin temperature based measures

Facial skin temperature can be employed as a type of non-intrusive, non-obtrusive, and real-time physiological measure for mental workload assessment. It has received increasing attention in cognitive workload studies as the cost of thermal infrared camera decreased in recent years. Especially, the skin temperature drop of nose area with increased mental workload has been observed in a few studies.

Veltman and Vos (2005) examined the variation of subject's facial skin temperature in a continuous memory task with two difficulty levels. The experimental results demonstrated the correlation between nose skin temperature and changes in mental workload. To enhance the sensitivity and accuracy, the facial skin temperature could be integrated with other physiological measures for cognitive workload evaluation.

Or and Duffy (2007) also studied changes in facial skin temperature for automated mental workload assessment. In the experiment, subjects performed driving test under different traffic conditions (city/highway) in simulator or real vehicle. Mental arithmetic test was used as a secondary task. Both forehead temperature and nose temperature were monitored during the experiment. It was found that under all simulator test conditions, nose skin temperature dropped significantly after the driving. The dual task of driving and arithmetic resulted in a greater nose temperature drop than the driving only task. In addition, the experimental results exhibited a significant correlativity between the nose skin temperature and the subjective rating of mental workload. Comparing with the real driving task, the simulated driving task had a higher subjective rating and it was observed with a greater change of nose skin temperature.

Previous research work on facial skin temperature has revealed its correlation to the variation of mental workload. However, it has also been noticed that the skin temperature based measures may not achieve sufficient sensitivity, especially for complex tasks or practical applications. Consequently, the combination of skin temperature and various other measures has been proposed to improve the performance of workload assessment. Wang et al. (2007) presented a composite workload index using three video based physiological measures, facial skin temperature, eye blinks, and pupil dilation. All the measures could be unobtrusively captured in real-time for workload evaluation.

7 Noisy factors in workload measures

Although a number of studies exhibited empirical evidence that eye activity based physiological measures could be used as an effective indicator of mental workload increase, the measures may fail to evaluate workload under complicated situations. For example, pupil dilation could be influenced by experimental environment like illumination condition. In addition to the experiment involving background lightness (Pomplun and Sunkara, 2003), Kramer (1991) reported the failure of workload measure due to factors unrelated to the cognitive task, such as changes in ambient illumination or screen luminance, which might give rise to greater variation of pupil size. In an experiment on the effects of perceptual/central and physical demands on physiological measures (Backs et al., 1994), it was found that physiological measures would be more sensitive to physical demands than to perceptual/central demands. In another experiment study of Sternberg memory search task (Van Gervan et al., 2003), the analysis results also demonstrated effects of aging on pupillary response. Moreover, to evaluate the usability of eye tracking data for cognitive workload measurement, Pomplun and Sunkara (2003) studied the distortion of pupil size caused by eye movements. The pupil size observed by the eye tracking camera would be affected by the gaze angle of the user. The eye tracking system was calibrated based on neural network to correct the geometry distortion of pupillary response data.

Video based physiological measures can also be influenced by a variety of affective factors including anxiety, engagement, fatigue, and stress (Chen, 2006; Pavlidis et al., 2000; Prinzel et al., 1999). For example, eye blinks, heart rate variability, or electroencephalogram (EEG) could be used to evaluate engagement and fatigue as well (Heishman and Duric, 2007; Zhang et al., 2008). Genno et al. (1997) investigated the changes in facial skin temperature caused by subject's stress or fatigue during the task. In the experiment of a task inducing stress, the nose skin temperature exhibited significant drop when the task started or an unexpected emergency alarm took place. Moreover, the nose skin temperature dropped significantly as well in the experiment of another task inducing fatigue. Meanwhile, Puri et al. (2005) also exhibited the correlation between forehead temperature and emotional state through thermal imaging.

Although it would be ideal to find a general model of human cognitive workload, mental workload could be personal characteristics of each subject. Thomas et al. (2009) studied personalized mental workload for exercise intensity measure. In the experiment, ratio of non-blink to blink frames and pupil radius were detected for each participant during different exercise tasks. It was suggested that due to non-stationary and nonzero-state nature of human being system, mental workload should be modelled individually and adaptively.

Table 2. Cognitive task–physiological measure matrix

Task	BF	BI	BD	FF	FD	SD	SS	PD	PC	PS	IC	ST	HM
Air traffic control task (Ahlstrom and Friedman-Berg, 2006)	●		●			●		●					
Air traffic control task (Brookings et al., 1996)	●					●							
Air traffic control task (Di Stasi et al., 2010)						●	●						
Auditory two-back task (Guhe et al., 2005)	●							●					●
Cart driving and stationary bike exercise (Thomas et al., 2009)								●					
Cognitive task and visual search task (Recarte et al., 2008)	●							●					
Combat management task (Greef et al., 2009)					●	●	●	●					
Continuous memory task (Veltman and Vos, 2005)												●	
Division task and Sternberg memory search (Murata and Iwase, 1998)										●			
Driving task and auditory addition task (Tsai et al., 2007)	●		●	●	●			●					
Driving task and secondary task (Schwalm et al., 2008)											●		
Driving task and spoken task (Palinko et al., 2010)								●					
Driving task and verbal/spatial-imagery task (Zhang et al., 2004)								●					
Document editing, email classification, route planning (Bailey and Iqbal, 2008)									●				
Flight task and memory task (Veltman and Gaillard, 1998)		●	●										
Flight task with visual/instrument flight rule (Wilson, 2002)	●												
Gaze-controlled interaction task (Pomplun and Sunkara, 2003)								●					
Language, visuospatial, and executive processing (Just et al., 2003)								●					
Mental arithmetic, short-term memory, aural vigilance (Klingner et al., 2008)								●					
Motorbike riding task (Di Stasi et al., 2009)				●	●	●	●	●					
Ocular following and oral calculation (Nakayama and Shimizu, 2004)								●		●			

Table 2. Cognitive task–physiological measure matrix (continued)

Task	BF	BI	BD	FF	FD	SD	SS	PD	PC	PS	IC	ST	HM
Reading, reasoning, searching, and object manipulation (Iqbal et al., 2004)									●				
Simulated/real driving task and mental arithmetic task (Or and Duffy, 2007)												●	
Tracking task and mental arithmetic task (Ryu and Myung, 2005)		●											
Tracking task and mental arithmetic task (Wang et al., 2007)		●						●				●	
Video game (action-puzzle) task (Lin and Imamiya, 2006)				●	●			●					●
Visual backward masking task (Verney et al., 2001)								●					
Visual horizontal tracking and visual gauge monitoring (Neumann, 2002)	●	●											
Visual search of symbolic displays (Backs and Walrath, 1992)				●	●			●					
Visuospatial memory task (Van Orden et al., 2001)	●		●	●	●	●		●					

Physiological measures. BF: blink frequency, BI: blink interval/latency, BD: blink duration, FF: fixation frequency, FD: fixation duration, SD: saccade distance/extent, SS: saccade speed, PD: pupil diameter/dilation, PC: percentage change in pupil size, PS: power spectrum, IC: index of cognitive activity, ST: skin temperature, HM: head/hand movement.

8 Multimodal measures and data fusion

Although physiological measures have exhibited reliable sensitivity to the variation of mental efforts when operators experience different levels of task demands, it is generally agreed that no single physiological measure can comprehensively describe cognitive workload. For example, in an experiment of actual flight scenario (Hankins and Wilson, 1998), eye activity only showed sensitivity to workload during flight segments that were visually demanding, meanwhile heart rate and EEG respectively showed sensitivity during flight segments of instrument rule and those requiring mental calculation. The experimental results demonstrated the multiple physiological measures could provide unique and non-overlapping information about subject's mental workload.

As multitasking is common in human activities, different subtasks may have different effects on individual physiological measures. In terms of the multiple resource theory for cognitive workload, the processing resource indexed by one video based physiological measure could be different from those indexed by other types of physiological measures. Table 2 lists recent research work using physiological measures for workload evaluation in various cognitive tasks. Multiple workload measures, especially physiological measures, could provide a comprehensive picture of the processing demands during the task execution. To further increase the performance of cognitive workload assessment, it is reasonable to combine different video based measures and/or other physiological measures.

8.1 Multiple measures vs. single measure

The sensitivity of individual physiological measures to workload demands could be very different. For example, in the experiment of different flight tasks (Veltman and Gaillard, 1998), cognitive workload measures including heart period, blood pressure, respiration, and eye blinks were recorded during the task. Although all the measures showed the difference between rest and fight, only heart period was sensitive to all the difficulty levels in the tunnel fight task.

Lin and Imamiya (2006) studied composite physiological measure through integrating eye movement and hand movement for mental effort evaluation when subjects performed a video game task. Although single physiological measures could only distinguish between the low difficulty level and high difficulty level, the composite measure was able to detect the variation of mental efforts for all the difficulty levels of the game task.

Similarly, Ryu and Myung (2005) showed that in the experiment with a dual task of tracking and arithmetic, none of the three physiological measures, including alpha suppression of brain activity, eye blink interval, and heart rate variability was able to identify the variation of the mental workload for both tasks. The alpha suppression was sensitive to the mental workload for the arithmetic task, but not for the tracking task. On the contrary, the blink interval and heart variability revealed sensitivity to the workload for the tracking task, but not for the arithmetic task. Although no single measures revealed sufficient sensitivity, significant variation of mental workload was successfully detected for both tasks when all these measures were combined altogether.

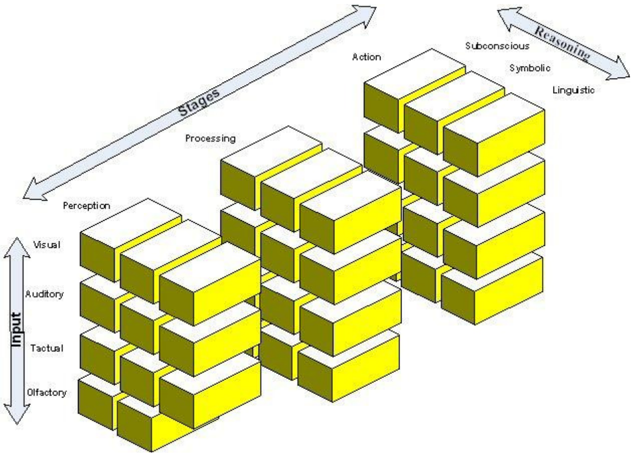


Figure 4. Multiple resource model (Wickens, 1984).

Consistent with the multiple resource theory (see Figure 4), previous studies indicated that task demands for different mental resource could be reflected by different physiological measures. The combination of multiple physiological measures has attracted increasing interests in cognitive workload studies, so that the explanatory power of multimodal information could be maximized.

8.2 Multimodal data fusion

The integration of multimodal information from multiple physiological measures is a non-trivial problem. Sometimes multiple measures could provide convergent results under single task condition, but inconsistent results under dual task condition. The way of data fusion is a key issue to efficiently and effectively integrate multimodal physiological features. For example, in a dual task experiment three workload measures based on brain activity, cardiac signal, and eye blink were combined into one composite measure using different weight coefficients (Ryu and Myung, 2005). It was shown that the composite measure significantly improved the sensitivity of workload assessment in the dual task.

Van Orden et al. (2001) employed artificial neural network to combine various eye activity based physiological features including blink frequency and duration, fixation frequency and time, saccadic extent, and pupil diameter for mental workload assessment. For each participant, a neural network model was trained on two sessions and tested on another session. Experimental results exhibited multiple eye activity based measures could be combined to produce reliable physiological index of workload in visuospatial task. In another experiment inducing fatigue (Van Orden et al., 2000), eye activity based features were also input to a neural network to estimate the fatigue state during the visual task performance.

Guhe et al. (2005) presented a Bayesian network approach to measure cognitive workload in real-time using multiple video based measures. The auditory two-back task, in which each participant was required to determine whether the current letter was equal to or different from the letter presented two back, was performed in the experiment. Video based features including blink frequency, eye closure, saccadic movement, eye gaze, pupil dilatation, head movement, and mouth openness were recorded for each participant in the experiment. To make the model adaptive to both individual users and the specific task, Bayesian network was employed to fuse multiple video based measures for mental workload evaluation.

Zhang et al. (2004) proposed a machine learning approach for driver workload estimation using multiple physiological features including eye gaze and pupil diameter. Instead of analysing the significance of individual measures, all the measures were considered simultaneously during the task. The estimation of cognitive workload was optimized automatically with the use of machine learning techniques such as decision tree and Bayesian learning.

The combination of eye activity based physiological measures and facial skin temperature has also been proposed to enhance the sensitivity of mental workload measurement. Wang et al. (2007) presented a composite workload index based on facial skin temperature, eye blinks and pupil dilation. To improve the overall sensitivity to cognitive workload, the way of integrating eye activity features and facial skin temperature would be constructed through factor analysis and regression analysis.

9 Future work

Besides its sensitivity to changes in mental workload and usability as an objective measure, video based physiological measure has an attractive advantage that the measurement data can be captured in a non-intrusive and non-obtrusive way. The imaging sensors, especially the remote ones, minimize user interference and enable continuous data acquisition. Therefore, it is expected that video based physiological measures will become more and more popular in research and application areas involving cognitive workload. Meanwhile, various technique issues could be further investigated to improve the overall accuracy and sensitivity for mental workload assessment.

Video based workload measures such as pupillary response and skin temperature may be influenced by noisy factors relating to sensor technology. For example, subtle changes in physiological measures could be ignored due to the insufficient accuracy or resolution of the sensor. For remote eye tracker, the pupil area observed in video frames is also affected by the pose of human face. The sensitivity of physiological measures could be further enhanced by correcting the noises and distortions introduced during the sensing process.

As cognitive workload is multidimensional, single dimension of workload may have different effects on individual physiological measures. Previous studies also showed that different physiological measures could provide both overlapping and non-overlapping information about cognitive workload. Hence it will be useful to study the correlation between various video based physiological measures, especially under multitasking conditions.

Multiple physiological features could provide more information and result in better evaluation of mental workload than single physiological input. However, simple combination methods such as voting or linear weighting might not improve the overall accuracy and sensitivity for cognitive workload assessment. With the development of machine learning and information fusion techniques in recent years, probabilistic models and tools such as dynamic Bayesian network and Markov decision process could be employed to improve the fusion of multiple physiological measures.

On the other hand, an operator's mental workload during a task is determined by both demands of the task and capacity of the subject. From previous work on mental workload measures, it has been observed that physiological data is sensitive to the levels of task difficulty. Besides, the physiological measures may exhibit the effects of cognitive capacity as well. The correlation between video based physiological measures and subject's capacity should be further investigated to improve the explanatory power of physiological data.

Furthermore, both cognitive workload and physiological measures are influenced by many factors. For example, cognitive workload is dependent on operator's level of training, expertise, experience, motivation, etc. On the other hand, physiological measures are affected by various factors such as fatigue, stress, engagement, and environment. Ignoring these aspects may lead to the failure of physiological measures for mental workload assessment. The efficiency and effectiveness of video based physiological measures could be significantly enhanced when more of these factors are considered in a comprehensive way.

10 References

[1] U. Ahlstrom, J. Friedman-Berg: Using eye movement activity as a correlate of cognitive workload. International Journal of Aviation Psychology, vol. 36, pp. 623–636, 2006.

[2] R. Backs, A. Ryan, G. Wilson: Psychophysiological measures during continuous and manual performance. Human Factors, vol. 36, pp. 514–531, 1994.

[3] R. Backs, L. Walrath: Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied Ergonomics, vol. 23, pp. 243–254, 1992.

[4] B. Bailey, S. Iqbal: Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management, ACM Transactions on Computer-Human Interaction, vol. 14, pp. 21-1–21-28, 2008.

[5] B. Bailey, C. Busbey, S. Iqbal: TAPRAV: An interactive analysis tool for exploring workload aligned to models of task execution, Interacting with Computers, vol. 19, pp. 314–329, 2007.

[6] J. Beatty: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, vol. 91, pp. 276–292, 1982.

[7] A. Boehm-Davis, D. Gray, J. Schoelles: The eye blink as a physiological indicator of cognitive workload, Human Factors and Ergonomics Society Annual Meeting, pp. 116–119, 2006.

[8] J. Brookings, G. Wilson, C. Swain: Psychophysiological responses to changes in workload during simulated air traffic control, Biological Psychology, vol. 42, pp. 361–377, 1996.

[9] A. Byrne, R. Parasuraman: Psychophysiology and adaptive automation. Biological Psychology, vol. 42, pp. 249–268, 1996.

[10] B. Cain, A review of the mental workload literature, Technical Report, Defence Research and Development Canada Toronto, 2007.

[11] J. Callicott et al.: Physiological characteristics of capacity constraints in working memory as revealed by functional MRI, Cerebral Cortex, vol. 9, pp. 20-26, 1999.

[12] F. Chen: Designing Human Interface in Speech Technology, pp. 53–94, Springer, 2006.

[13] J. Coyne, C. Baldwin, A. Cole, C. Sibley, D. Roberts: Applying real time physiological measures of cognitive load to improve training, International Conference on Human-Computer Interaction, pp. 469–478, 2009.

[14] L. Di Stasi, V. Álvarez-Valbuena, J. Cañas, A. Maldonado, A. Catena, A. Antolí, A. Candido: Risk behaviour and mental workload: Multimodal assessment techniques applied to motorbike riding simulation, Transportation Research Part F: Traffic Psychology and Behaviour, vol. 12, pp. 361–370, 2009.

[15] L. Di Stasi, M. Marchitto, A. Antolí, T. Baccino, J. Cañas, Approximation of on-line mental workload index in ATC simulated multitasks, Journal of Air Transport Management, in press, 2010.

[16] C. Fogarty, J. Stern: Eye movements and blinks: Their relationship to higher cognitive processes. International Journal of Psychophysiology, vol. 8, pp. 35–42, 1989.

[17] L. Fournier, G. Wilson, C. Swain: Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training, International Journal of Psychophysiology, vol. 31, pp. 129–145, 1999.

[18] H. Genno, K. Ishikawa, O. Kanbara, M. Kikumoto, Y. Fujiwara, R. Suzuki, M. Osumi: Using facial skin temperature to objectively evaluate sensations. International Journal of Industrial Ergonomics, vol. 19, pp. 161–171, 1997.

[19] R. Gingell, Review of workload measurement, analysis and interpretation methods, Technical Report, European Organisation for the Safety of Air Navigation, 2003.

[20] D. Gopher, E. Donchin: Workload-An examination of the concept. K. Boff, L. Kaufman, J. Thomas, (eds.) Handbook of Perception and Human Performance, Wiley, 1986.

[21] T. Greef, H. Lafeber, H. Oostendorp, J. Lindenberg: Eye movement as indicators of mental workload to trigger adaptive automation, International Conference on Human-Computer Interaction, pp. 219–228, 2009.

[22] M. Grootjen, M. Neerincx, J. Weert: Task-based interpretation of operator state information for adaptive support. D. Schmorrow, M. Stanney, M. Reeves, (eds.) Foundations of Augmented Cognition (2nd edn.), pp. 236–242, 2006.

[23] M. Grootjen, M. Neerincx, J. Weert, and K. Truong: Measuring cognitive task load on a naval ship: implications of a real world environment, International Conference on Human-Computer Interaction, pp. 147–156, 2007.

[24] M. Guhe, W. Liao, Z. Zhu, Q. Ji, D. Gray, J. Schoelles: Non-intrusive measurement of workload in real-time. Human Factors and Ergonomics Society Annual Meeting, pp. 1157–1161, 2005.

[25] T. Hankins, G. Wilson: A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviation, Space, and Environmental Medicine, vol. 69, pp. 360–367, 1998.

[26] S. Hart, L. Staveland: Development of the NASA task load index (TLX): Results of experimental and theoretical research. P. Hancock, N. Meshkati (eds.): Human Workload, pp. 138–183, North-Holland, 1988.

[27] P. He, B. Yang, S. Hubbard, J. Estepp, G. Wilson: A sensor positioning system for functional near-infrared neuroimaging. International Conference on Human-Computer Interaction, pp. 30–37, 2007.

[28] R. Heishman, Z. Duric: Using eye blinks as a tool for augmented cognition. International Conference on Human-Computer Interaction, pp. 84–93, 2007.

[29] S. Iqbal, X. Zheng, B. Bailey: Task-evoked pupillary response to mental workload in human-computer interaction. ACM Conference on Human Factors in Computing Systems (CHI), pp. 1477–1480, 2004.

[30] M. Izzetoglu, K. Izzetoglu, S. Bunce, H. Ayaz, A. Devaraj, B. Onaral, K. Pourrezaei: Functional near-infrared neuroimaging. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 13, pp. 153–159, 2005.

[31] M. Just, P. Carpenter, A. Miyake: Neuroindices of cognitive workload neuroimaging, pupillometric and event-related potential studies of brain work, Theoretical Issues in Ergonomics Science, vol. 4, pp. 56–88, 2003.

[32] J. Klingner, R. Kumar, P. Hanrahan: Measuring the task-evoked pupillary response with a remote eye tracker. Eye Tracking Research and Applications Symposium, pp. 69–72, 2008.

[33] A. Kramer: Physiological metrics of mental workload: A review of recent progress, D. Damos (ed.), Multiple-Task Performance, pp. 279–328, Taylor and Francis, 1991.

[34] T. Lin, A. Imamiya: Evaluating usability based on multimodal information: An empirical study. International Conference on Multimodal Interfaces, pp. 364–371, 2006.

[35] P. Marshall: The index of cognitive activity: Measuring cognitive workload. IEEE Human Factors Meeting, pp. 7-5–7-9, 2002.

[36] J. May, R. Kennedy, M. Williams, W. Dunlap, J. Brannan: Eye movement indices of mental workload. Acta Psychologica, vol. 75, pp. 75–89, 1990.

[37] A. Murata, H. Iwase: Evaluation of mental workload by fluctuation analysis of pupil area. Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 3094–3097, 1998.

[38] M. Nakayama, Y. Shimizu: Frequency analysis of task evoked pupillary response and eye-movement, Eye Tracking Research and Applications Symposium, pp. 71–76, 2004.

[39] D. Neumann: Effect of varying levels of mental workload on startle eyeblink modulation. Ergonomics, vol. 45, pp. 583–602, 2002.

[40] R. O'Donnell, F. Eggemeier: Workload assessment methodology. K. Boff, L. Kaufman, J. Thomas (eds.): Handbook of Perception and Human Performance. Cognitive Processes and Performance, vol. 2, pp. 42-1–42-49, Wiley, 1986.

[41] K. Or, G. Duffy: Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. Occupational Ergonomics, vol. 7, pp. 83–94, 2007.

[42] F. Paas, E. Juhani, H. Tabbers, P. Van Gerven: Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist, vol. 38, pp. 63–71, 2003.

[43] O. Palinko, A. Kun, A. Shyrokov, P. Heeman: Estimating cognitive load using remote eye tracking in a driving simulator, Eye Tracking Research and Applications Symposium, pp. 141–144, 2010.

[44] I. Pavlidis, J. Levine, P. Baukol: Thermal imaging for anxiety detection. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, pp. 104–109, 2000.

[45] M. Pomplun, S. Sunkara: Pupil dilation as an indicator of cognitive workload in human-computer interaction. International Conference on Human-Computer Interaction, pp. 542–546, 2003.

[46] L. Prinzel, F. Freeman, M. Scerbo, P. Mikulka, A. Pope: A closed-loop system for examining psychophysiological measures for adaptive task allocation. International Journal of Aviation Psychology, pp. 393–410, 1999.

[47] C. Puri, L. Olson, I. Pavlidis, J. Levine, J. Starren: StressCam: Non-contact measurement of users' emotional states through thermal imaging. ACM conference on Human factors in computing systems (CHI), pp. 1725–1728, 2005.

[48] M. Recarte, E. Perez, A. Conchillo, L. Nunes: Mental workload and visual impairment: differences between pupil, blink, and subjective rating. Spanish Journal of Psychology, vol. 11, pp. 374–385, 2008.

[49] K. Ryu, R. Myung: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. International Journal of Industrial Ergonomics, vol. 35, pp. 991–1009, 2005.

[50] M. Scerbo, F. Freeman, P. Mikula, R. Parasuraman, F. Di Nocero, L. Prinzel: The efficacy of psychophysiological measures for implementing adaptive technology. Technical Report, NASA Langley Research Center, 2001.

[51] M. Schwalm, A. Keinath, H. Zimmer: Pupillometry as a method for measuring mental workload with a simulated driving task, D. Waard, F. Flemisch, B. Lorenz, H. Oberheid, K. Brookhuis (eds.), Human Factors for Assistance and Automation, pp. 75–87, 2008.

[52] L. Sciarini, D. Nicholson: Assessing cognitive state with multiple physiological measures: A modular approach, International Conference on Human-Computer Interaction, pp. 533–542, 2009.

[53] J. Sirevaag, A. Stern: Ocular measures of fatigue and cognitive factors. W. Backs, W. Boucsein (eds.): Engineering Psychophysiology - Issues and applications, pp. 269–287, 2000.

[54] J. Sweller: Cognitive load during problem solving: Effects on learning. Cognitive Science: A Multidisciplinary Journal, vol. 12, pp. 257–285, 1988.

[55] N. Thomas, Y. Du, T. Artavatkun, J. She: Non-intrusive personalized mental workload evaluation for exercise intensity measure, International Conference on Human-Computer Interaction, pp. 315–322, 2009.

[56] Y. Tsai, E. Viirre, C. Strychacz, B. Chase, T. Jung: Task performance and eye activity - predicting behavior relating to cognitive workload. Aviation, Space, and Environmental Medicine, vol. 78, pp. B176–B185, 2007.

[57] M. Tungare, M. Perez-Quinones: Mental workload in multi-device personal information management, ACM Conference on Human Factors in Computing Systems (CHI), pp. 3431–3436, 2009.

[58] P. Van Gerven, F. Paas, J. Van Merriënboer, H. Schmidt: Memory load and the cognitive pupillary response in aging, Psychophysiology, vol. 41, pp. 167–174, 2003.

[59] K. Van Orden, T. Jung, S. Makeig: Combined eye activity measures accurately estimate changes in sustained visual task performance. Biological Psychology, vol. 52, pp. 221–240, 2000.

[60] K. Van Orden, W. Limbert, S. Makeig, T. Jung.: Eye activity correlates of workload during visuospatial memory task. Human factors, vol. 43, pp. 111–121, 2001.

[61] H. Veltman, W. Vos: Facial temperature as a measure of operator state. International Conference on Augmented Cognition, pp. 22–27, 2005.

[62] J. Veltman, A. Gaillard: Physiological workload reactions to increasing levels of task difficulty. Ergonomics, vol. 41, pp. 656–669, 1998.

[63] J. Veltman, C. Jansen: The role of operator state assessment in adaptive automation, Technical Report, TNO Human Factors Research Institute, 2006.

[64] S. Verney, E. Granholm, D. Dionisio: Pupillary responses and processing resources on the visual backward masking task. Psychophysiology, vol. 38, pp. 76–83, 2001.

[65] J. Voskamp, B. Urban: Measuring cognitive workload in non-military scenarios criteria for sensor technologies, International Conference on Human-Computer Interaction, pp. 304–310, 2009.

[66] L.-M. Wang, V. Duffy, Y. Du: A composite measure for the evaluation of mental workload, International Conference on Human-Computer Interaction, pp. 460–466, 2007.

[67] C. Wickens: Processing resources in attention. R. Parasuraman, D. Davies, (eds.) Varieties of Attention, pp. 63–102. Academic Press, 1984.

[68] C. Wickens: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, vol. 3, pp. 150–177, 2002.

[69] W. Wierwille, F. Eggemeier: Recommendations for mental workload measurements in a test and evaluation environment. Human Factors, vol. 35, pp. 263–281, 1993.

[70] G. Wilson: An analysis of mental workload in pilots during flight using multiple psychophysiological measures. International Journal of Aviation Psychology, vol. 12, pp. 3–18, 2002.

[71] G. Wilson et al.: Operator functional state assessment. Technical Report, Research and Technology Organisation, North Atlantic Treaty Organisation (NATO), 2004.

[72] G. Wilson, C. Russel: Psychophysiologically versus task determined adaptive aiding accomplishment. D. Schmorrow, K. Stanney, L. Reeves (eds.): Foundations of Augmented Cognition (2nd edn.), pp. 201–207, 2006.

[73] B. Yin, F. Chen, N. Ruiz, E. Ambikairajah: Speech-based Cognitive Load Monitoring System, IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 2041–2044, 2008.

[74] C. Zhang, C. Zheng, X. Yu: Evaluation of mental fatigue based on multipsychophysiological parameters and kernel learning algorithms, Chinese Science Bulletin, vol. 53, pp. 1835–1847, 2008.

[75] Y. Zhang, Y. Owechko, J. Zhang: Driver cognitive workload estimation - a data driven perspective. International IEEE Conference on Intelligent Transportation Systems, pp. 642–647, 2004.

Appendix B

An Investigation of Formant Frequencies for Cognitive Load Classification

Tet Fei Yap[1,2], *Julien Epps*[1,2], *Eliathamby Ambikairajah*[1,2], *Eric H. C. Choi*[2]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia
tetfei.yap@nicta.com.au, j.epps@unsw.edu.au, ambi@ee.unsw.edu.au, eric.choi@nicta.com.au

Abstract

The cognitive load experienced by a person can be used as an index to monitor task performance. Hence, the ability to measure the cognitive load of a person using speech can potentially be very useful, especially in areas such as air traffic control systems. Current research on cognitive load does not provide enough insight into how cognitive load affects the speech spectrum, or the speech production system. Since formants are closely related to the underlying vocal tract configuration, this work aims to study the effect of cognitive load on vowel formant frequencies, and hence proposes the effective application of formant features to cognitive load classification. Results from classification performed on the Stroop test database show that formant features not only have lower dimensionality, but dynamic formant features can outperform conventionally used MFCC-based features by a relative improvement of 12%.

Index Terms: formant frequency, cognitive load classification, Gaussian mixture models

1. Introduction

Cognitive load refers to the load imposed by a task on the cognitive system of a person [1]. It is related to the limited amount of human working memory. A difficult task typically requires a larger amount of working memory to complete; it is said to induce a higher cognitive load. Cognitive load theory plays an important role in systems wherein the task performance is significantly affected by the cognitive load experienced; for example, air traffic control systems. In such situations, it is often desirable to measure and monitor the cognitive load experienced.

Different methods have been proposed to measure cognitive load [1], but speech is an attractive option as it is non-intrusive and its processing can potentially be real-time. Research that correlates speech parameters with workload has been conducted for quite some time [2], but it was only in 2008 that Yin *et al.* proposed a fully automatic speech-based cognitive load classification system [3].

Since then, research has focused on frame-based acoustic features for Gaussian mixture model (GMM)-based classifiers [4, 5], and Mel-frequency cepstral coefficients (MFCCs) have emerged as effective features for cognitive load classification. However, MFCCs are ill-suited to answer the questions of how cognitive load affects the speech spectrum, or the underlying physical speech production system; the answers are important as they provide insight and motivation for the development of better features for classification.

Previously, we have investigated features related to the speech production system; i.e., glottal features [5]. A study of formant frequencies is, hence, relevant since formants are closely related to the physical characteristics of the vocal tract. Although Lively *et al.* found no observable changes in the first three formant frequencies under different workload conditions [2], other studies have shown that formants are affected by emotion dimensions [6], and can be used for stress classification [7]. Our previous investigations have also shown formant frequencies to be effective features for cognitive load classification [8].

In general, studies of speech under different cognitive load have focused mainly on testing the correlation of speech parameters and cognitive load [2], or deriving features for classification purposes [3, 4, 8]. To the best of our knowledge, there has yet to be a study that considers both aspects at the same time. Moreover, our previous work [8] did not examine the effect of cognitive load at the level of individual vowels. Hence, in this paper, we provide a detailed analysis of the vowel-level effect of cognitive load on formants, together with classification results for different formant feature combinations.

2. Analysis of vowel formant frequencies under cognitive load

2.1. Cognitive load database

The database used in this work is the Stroop test database [3], which consists of speech recorded from 16 randomly selected native English speakers (7 male and 9 female) performing three tasks of varying cognitive load. In the low load task, speakers were asked to read aloud words corresponding to color names. In the medium load task, a mismatch was introduced between the color names and their font colors, and the speakers were asked to name the font colors instead. The high load task was similar to the medium load task except that time constraints were added to the task. Approximately 54 recordings were obtained per load level; in each recording, 20 color names were uttered (10 different colors randomly repeated twice). A reading task of about 90s was also recorded by the same participants.

2.2. Experimental setup

A subset of the Stroop test database, comprising 4 vowel sounds (/ae/, /eh/, /iy/ and /uw/ extracted from the words 'black', 'red', 'green' and 'blue' respectively), was selected for this set of experiments. The vowel sounds were chosen to be well-separated in the vowel space: /ae/, /iy/ and /uw/ are typically situated at different corners of the F1-F2 vowel plane.

Only vowels spoken under low and medium cognitive load conditions were analyzed. This is because the high load task design induces a change in rate of speech. Since the effect of speech rate on formant frequencies has been well studied [9],

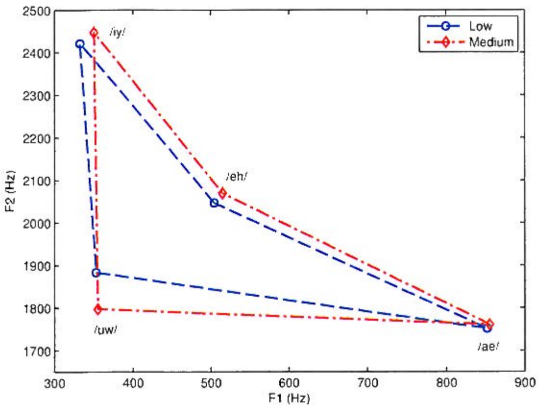


Figure 1: Means of the first two formant frequencies for 4 different vowel sounds, averaged across all speakers.

we focused only on the low and medium load tasks, wherein changes in speech rate are not explicitly induced.

The vowel sounds were extracted from color words, manually segmented from full utterances in the Stroop test database. The phone boundaries of the vowel sounds were then obtained by performing forced phone alignment on the color words, using the Hidden Markov Model Toolkit (HTK). The phone models were trained, using MFCC features, on the training partition of the TIMIT corpus [10]. Performance of this phone alignment system, when tested on the test partition of the TIMIT corpus, was 87.2% with a temporal tolerance of 20ms.

The first three formant frequencies were then extracted from each vowel sound using the Wavesurfer/Snack toolkit [11]. A 49ms Hamming window was applied with a frame increment of 10ms. Apart from that, all other parameters remained at the default settings: the number of formants tracked (four), pre-emphasis factor (0.7), LPC order (12) and sampling frequency (10kHz). Erroneous formant values were manually corrected using a spectrogram.

Speaker-specific feature warping [8, 12] was applied to the formant frequencies when formant frequencies across all speakers were analyzed with a GMM (Section 2.3.2). This was to remove inter-speaker variability, such as differences in vocal tract lengths among speakers. This method seeks to map the probability distribution of the pooled formant frequencies of each speaker to the standard normal distribution. For a given formant value p for speaker S, the warped feature value q is given by the equation $q = H^{-1}\left(\frac{N+\frac{1}{2}-R}{N}\right)$, where $H^{-1}()$ denotes the normal inverse cumulative distribution function, N is the total number of formant values for speaker S and R is the ranking of p after sorting the formant values for speaker S in descending order. A detailed explanation can be found in [12].

2.3. Results

2.3.1. Effect of cognitive load on vowel formant means

Figure 1 shows the mean F1 and F2 formant frequencies, averaged across all speakers, for the 4 different vowel sounds spoken under low or medium cognitive load. This figure suggests that vowel formant frequencies are shifted under different cognitive load conditions, and these shifts are vowel-dependent.

In order to better visualize the direction of the formant shift, we plotted the speaker averaged formant frequency difference $\bar{F}^{(v)}_{i,\Delta} = \frac{1}{N}\sum_{k=1}^{N}\left(\bar{F}^{(v)}_{i,k,med} - \bar{F}^{(v)}_{i,k,low}\right)$, where $\bar{F}^{(v)}_{i,k,low}$ and $\bar{F}^{(v)}_{i,k,med}$ are the i-th formant frequencies under low and medium cognitive load respectively, spoken by speaker k, and averaged across all instances of vowel sound v. A positive (or negative) value of $\bar{F}^{(v)}_{i,\Delta}$ indicates an increasing (or decreasing) trend for the i-th formant frequency, whereas a value of zero indicates no change from low to medium cognitive load conditions.

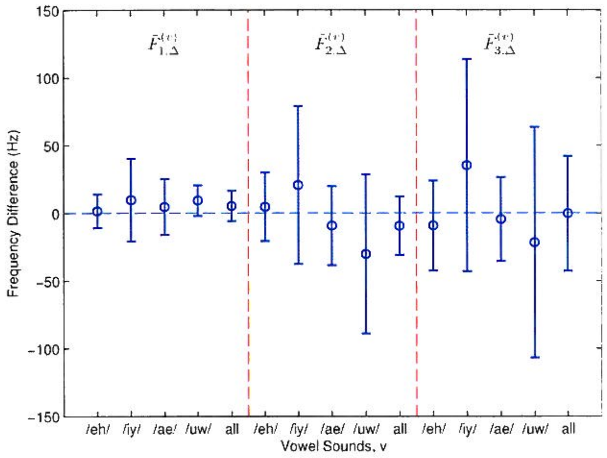


Figure 2: Formant frequency difference $\bar{F}^{(v)}_{i,\Delta}$ from low to medium cognitive load conditions for different vowel sounds.

Figure 2 shows the values of $\bar{F}^{(v)}_{i,\Delta}$ for different vowel sounds, together with their 95% confidence intervals. Clearly, the confidence intervals are quite wide and they overlap at zero mean. This is unsurprising considering that the analysis is performed on a small database. The confidence intervals seem to be larger for higher order formants, but this is most likely due to the increasing mean values across F1, F2 and F3. Formant frequencies are sometimes plotted on a log scale; when this was done, the confidence intervals obtained were approximately the same size.

Looking at just the mean values of the differences, F1 exhibits a generally increasing trend as cognitive load is increased, most notably for /iy/ and /uw/. F2, on the other hand, shows both an increasing trend (for /iy/ and /eh/) and a decreasing trend (/ae/ and /uw/). F3 exhibits a decreasing trend for most vowel sounds, except /iy/.

By way of comparison with other formant trends reported in the literature, we also looked at the formant frequency difference $\bar{F}^{(v)}_{i,\Delta}$ averaged across all 4 vowel sounds. This represents the average effect of cognitive load on the vowel formant frequencies. Results shown in Figure 2 suggest that when all 4 vowels are considered together, as cognitive load increases from low to medium, F1 is increasing and F2 is decreasing, whereas F3 remains unchanged. A larger database is necessary to validate these findings. Nevertheless, the consistent increase in F1 agrees with the findings of Hansen *et al.*, wherein F1 increases when speech is produced under stressful conditions [7].

2.3.2. GMM-based vowel formant analysis

One limitation of the analysis discussed in the previous section is that the vowel formant shifts were studied using the mean formant values alone. In a practical (automatic) cognitive load classification system, a frame-based approach is used instead. Also, differences across speakers are reduced through normalization techniques. Hence, it is desirable to discuss the

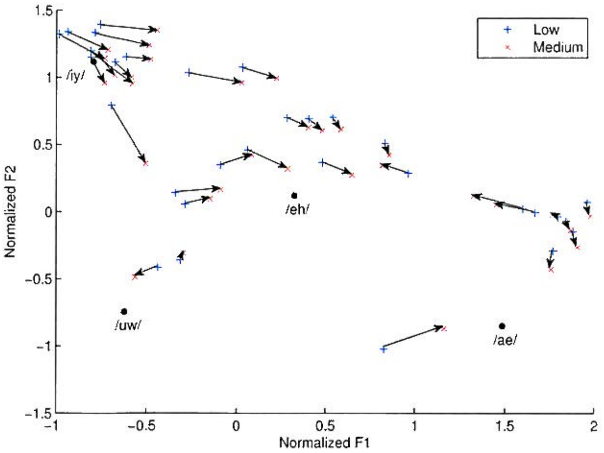


Figure 3: Movement of Gaussian mixture means (trained using $\{F_1, F_2\}$ features across all speakers) from low to medium cognitive load. (Filled circle represents mean formant frequencies for a particular vowel, averaged across all speakers).

formant shifts from the perspective of a practical classification system. Since a GMM classifier has, to date, been the standard back-end system employed in cognitive load classification systems [3, 4, 5, 8], we decided to study the vowel formant shifts using a GMM-based approach as well.

A universal background model (UBM), with 32 Gaussian mixtures, was trained using pairs of uncorrected but normalized formant frequencies, $\{F_1, F_2\}$ or $\{F_1, F_3\}$, extracted from the reading task data of all speakers in the Stroop test database. Maximum a-posteriori (MAP) adaptation was then applied to the means of the UBM using manually corrected and normalized $\{F_1, F_2\}$ or $\{F_1, F_3\}$ extracted from the 4 vowel sounds of all speakers, under different load conditions.

Figure 3 shows the shifts in Gaussian mixture means under low and medium cognitive load conditions, when F_1 and F_2 were considered. The effect of cognitive load on different vowel regions can be analyzed from this plot, since the 4 vowel sounds were chosen to be reasonably separated in the vowel space. The arrows in the plot indicate the movements of the Gaussian mixture means from low to medium load.

Overall, majority of the arrows in Figure 3 indicate that F_1 is increasing, while F_2 is decreasing, as cognitive load is increased. These trends are clearest in the /eh/ and /iy/ vowel regions. As for F_3, although the plots are not shown in this paper, we observed an overall decreasing trend. It is worth noting that increasing F1 and decreasing F2 trends were also observed when formant frequency differences, averaged across all vowel sounds, were considered in Section 2.3.1 (Figure 2), although these must be considered as preliminary results.

Given the observations made in this subsection, we expect formant features to perform well in a GMM-based cognitive load classification system. This is not only due to the general shifts in formant frequencies for different cognitive load, but also due to the separation of the different vowels in the acoustic space, and the ability of the GMM to model the changes in the distributions of different vowel regions.

2.3.3. *Effect of cognitive load on vowel formant trajectories*

Figure 4 shows the mean trajectories of the vowel sounds in the F1-F2 plane, averaged across all speakers. The trajectories were time-aligned using linear interpolation. From the figure, formant trajectories for most vowel sounds are similar across different cognitive load. However, the vowel sound /uw/ shows distinctly different trajectories under different cognitive load. This was verified by checking the spectrogram and also listening to the individual vowel sounds. The differences in shapes observed here suggest the need to model the formant trajectory in order to capture the formant dynamic information. In this paper, dynamic formant information was captured using regression-based delta coefficients.

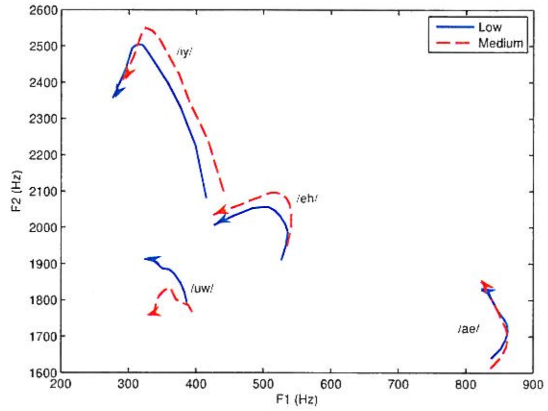


Figure 4: F1-F2 formant trajectories of different vowel sounds for low and medium cognitive load conditions.

3. Cognitive load classification using formant features

3.1. Experimental setup

The effectiveness of formant features for automatic cognitive load classification was tested by performing a leave-one-speaker-out cross validation on all utterances in the Stroop test database, rather than just the vowel subset used in Section 2.

Unlike the analysis in Section 2, all 3 load levels were considered in this evaluation to facilitate comparisons with the classification results presented in other work [3, 8]. It is worth noting that almost all of the experimental findings for the 3-class evaluation below also held in the 2-class case.

As with previous classification systems [3], a GMM classifier was used as the back-end system. A universal background model, with 32 Gaussian mixtures, was trained using speech data obtained from the reading task. MAP adaptation was then performed to adapt the means of the UBM using speech data obtained from the Stroop test task.

Features that were considered here include the various combinations of normalized formant frequencies extracted using the method outlined in Section 2.2, with the exception that erroneous formant values were not manually corrected. Dynamic formant information was also considered through the use of regression-based delta coefficients. A regression window size of 9 frames was found to provide reasonably good classification performance.

The individual formant features were combined using two different methods. Feature level fusion involved simple concatenation of the individual formant features. Score level fusion, on the other hand, involved combining the loglikelihood scores from two systems utilizing different features [5]. Given the i-th class loglikelihood scores $L_{1,i}$ and $L_{2,i}$ returned by two separate GMM systems, the fused loglikelihood scores can then be calculated as $L_{fused,i} = \alpha_i L_{1,i} + (1 - \alpha_i) L_{2,i}$, where

Table 1: 3-class cognitive load classification results using formant and MFCC features.

	Accuracy (%)	
Feature	Without Delta	With Delta
$MFCC$	55.8	60.2
$\{F_1, F_2, F_3\}$	55.2	67.7
$\{F_1, F_2\}$	55.8	65.2
$\{F_1, F_3\}$	51.0	64.5
$\{F_2, F_3\}$	44.7	60.3
F_1	55.4	60.9
F_2	53.4	58.9
F_3	44.8	44.1
Score Level Fusion		
Feature	Without Delta	With Delta
$F_1 + F_2$	61.5	67.9
$F_2 + F_3$	57.7	57.7
$F_1 + F_3$	50.0	58.3

$\alpha_i \in [0, 1]$ is the fusion weight of the i-th class. The value of α_i for a particular fold was determined by finding the optimum weights using the loglikelihood scores from all other folds. The cognitive load level decision was then determined from L_{fused}.

The classification performance of the formant features was then compared with that of the commonly used 7 MFCCs (with the zeroth order coefficient dropped) [3], with and without delta coefficients.

3.2. Results

The classification results, presented in Table 1, imply that the normalized formant frequencies, F_1, F_2 and F_3, are all capable of discriminating between different cognitive load levels, since their classification performance were all well above chance level.

The first two formant frequencies, F_1 and F_2, seem to be more effective than F_3. The good performance of F_1 is consistent with the findings of Le *et al.*, wherein most of the important cognitive load information derived from static spectral magnitude-based features was shown to be in the frequency region below 1kHz [4]. One possible reason for the relatively lower classification performance of F_3 is the inherent difficulty of automatically extracting formants of higher order. In fact, using $\{F_1, F_2\}$ as features was just as effective as using $\{F_1, F_2, F_3\}$.

When static formant features were considered, the best classification performance (61.5%) was obtained by performing score level fusion on F_1 and F_2 systems. This suggests that F_1 and F_2 carry complementary cognitive load information. In terms of dynamic formant features, classification performance consistently improved when delta coefficients were added to the formant features. This agrees with previous findings that dynamic information is important for cognitive load classification [3, 8]. Once again, the best classification performance (67.9%) was obtained by performing score level fusion on $\{F_1, \Delta F_1\}$ and $\{F_2, \Delta F_2\}$ systems, although $\{F_1, F_2, F_3, \Delta F_1, \Delta F_2, \Delta F_3\}$ produced an approximately similar result (67.7%).

When the formant features $\{F_1, F_2, F_3\}$ were compared with MFCC features, results showed that formant features not only have lower dimensionality, but provided comparable if not better performance. When static features were considered, the classification performance of formant features (number of dimensions, dim = 3) was similar to that of MFCCs (dim = 7). When dynamic delta features were incorporated into the feature sets, formant features (dim = 6) outperformed MFCCs (dim = 14) by a relative improvement of 12%.

4. Conclusion

This paper has studied formant frequencies from two different perspectives: firstly, from the point of view of trying to understand the effect of cognitive load on formants, and hence the speech production system, and secondly, from the point of view of effective features for cognitive load classification. In general, F1 was found to be increasing, and F2 decreasing, when cognitive load was increased. Additionally, we have also found that formant frequencies exhibited vowel-specific shifts in their mean values under different cognitive load conditions. For classification purposes, not only did formant features have lower dimensionality compared with MFCCs, they also outperformed MFCCs by a relative improvement of 12% when dynamic information was incorporated in both feature sets. Future work will, hence, include an investigation of different features to capture dynamic formant information, and of the use of these features for improved cognitive load classification.

5. References

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, Mar. 2003.

[2] S. E. Lively, D. B. Pisoni, W. V. Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2962–2973, 1993.

[3] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *Proceedings of ICASSP*, 2008, pp. 2041–2044.

[4] P. N. Le, E. Ambikairajah, E. H. C. Choi, and J. Epps, "A non-uniform subband approach to speech-based cognitive load classification," in *Proceedings of ICICS*, 2009, pp. 1–5, to appear.

[5] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification," in *Proceedings of ICASSP*, 2010, pp. 5234–5237.

[6] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Proceedings of Interspeech*, 2009, pp. 1575–1578.

[7] J. H. L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," *Lecture Notes in Computer Science*, vol. 4343, pp. 108–137, Aug. 2007.

[8] T. F. Yap, E. Ambikairajah, J. Epps, and E. H. C. Choi, "Cognitive load classification using formant features," in *Proceedings of ISSPA*, 2010, pp. 221–224.

[9] M. Pitermann, "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3425–3437, Jun. 2000.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-100354*, 1993.

[11] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of International Conference on Spoken Language Processing*, vol. 4, 2000, pp. 464–467.

[12] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker normalisation for speech-based emotion detection," in *Proceedings of International Conference on Digital Signal Processing*, Jul. 2007, pp. 611–614.

Appendix C

Eye Activity as a Measure of Human Mental Effort in HCI

Siyuan Chen[1,2] **Julien Epps**[1,2]
[1] Electrical Engineering and Telecommunications
The University of New South Wales
Kensington, NSW, Australia
siyuan.chen@student.unsw.edu.au
j.epps@unsw.edu.au

Natalie Ruiz[2] **Fang Chen**[2]
[2]NICTA
Level 5, 13 Garden street, Eveleigh
NSW, Australia
natalie.ruiz@nicta.com.au
fang.chen@nicta.com.au

ABSTRACT

The measurement of a user's mental effort is a problem whose solutions may have important applications to adaptive interfaces and interface evaluation. Previous studies have empirically shown links between eye activity and mental effort; however these have usually investigated only one class of eye activity on tasks atypical of HCI. This paper reports on research into eight eye activity based features, spanning eye blink, pupillary response and eye movement information, for real time mental effort measurement. Results from an experiment conducted using a computer-based training system show that the three classes of eye features are capable of discriminating different cognitive load levels. Correlation analysis between various pairs of features suggests that significant improvements in discriminating different effort levels can be made by combining multiple features. This shows an initial step towards a real-time cognitive load measurement system in human-computer interaction.

Author Keywords

Mental effort, cognitive load, pupillary response, eye blink, eye movement, adaptive interfaces.

ACM Classification Keywords

H5.2. User Interfaces: *Evaluation/methodology*.

General Terms

Measurement, human factors.

INTRODUCTION

Recently, research effort has been devoted towards making human-computer interaction closer to human-human communication, exemplified by affective computing for emotional intelligence, animated interface agents for trustworthiness, adaptive systems for personalized access and the development of multimodal user interfaces for increased flexibility. During interaction, an important characteristic of users is that their limited attention and working memory affect the commitment of cognitive resources, and users' different cognitive strengths and limitations affect individual task performance. Quantifying and monitoring human mental effort, therefore, holds the potential to prevent cognitive overload and provides support for a streamlined interaction without task failure.

With the aim of measuring mental effort, performance scoring (e.g. reaction time and accuracy) is an alternative recourse, likewise the subjective self-rating of users' perceptions of their tasks can be employed; however these rely on overt and discrete responses, and above all they are post-processing measurements [2]. Mental effort, on the other hand, is another aspect of cognitive load and is associated with cognitive capacity. It is believed to reflect cognitive load and is embodied by physiological variables [4]. Recent ubiquitous computing technology greatly facilitates the application of physiological techniques, through portability, unobtrusiveness and real time measurement. In this paper, our aim is to take the first step in developing a real time cognitive load measurement, based on one such physiological technique as an effective means of measuring how much mental effort has been devoted and in turn, whether the cognitive load limit has been reached.

RELATED WORK

Among all possible physiological measures, eye activity provides rich information about cognition and human mental effort. For example, task-invoked pupillary response is believed to be mainly due to the decrease in parasympathetic activity in the peripheral nervous system and has been found to vary linearly with the amount of information processed in short-term and long-term memory tasks as well as task difficulty levels [2]. Endogenous eye blinks are controlled by the central nervous system and tend to be inhibited during attention-demanding tasks to maximize stimulus perception [1,2]. Fixations and saccades are the main forms of the central controlled eye movement, which is thought to be a combination of bottom-up and top-down processes. The first sweep of the scene is basically a feature-driven process and more top-down control is varied by the effort to spread attention across the visual field to selected task-related objects [3].

Previous research on correlates of eye activity has mainly focused on single measures as indicators of workload in specific tasks such as working memory span [7] or attention-demanding tasks (for a general review see [1,2]).

IUI 2011, February 13–16, 2011, Palo Alto, California, USA.
Copyright 2011 ACM 978-1-4503-0419-1/11/02...$10.00.

Some used two or three classes of eye activity in simulated warfare management tasks [5,6]. In this paper, we measure human mental effort through the three classes of eye activity collected in a computer-based training system. We examined their sensitivity for analysis of user mental effort, with implications for user modeling and interface design. This study, in a semi-realistic scenario, is one of multiple steps towards a real-time system to measure human mental effort via eye activity.

RESEARCH METHOD

Participants

Twelve paid male recreational basketball players, each with more than two years' experience, aged 19-36, completed this experiment.

Task Description and Procedure

A computer-based training application, running on a tablet monitor, was designed for basketball players to learn playing strategies by observing team player positions in basketball game videos. The goal of this task was to detect and identify defenders and attackers during a video clip of an actual game, and recall their positions around the ball at the end of each 15-second clip.

During each session, subjects were seated in a quiet room. A head banded eye tracker was then attached. The camera angle was fixed during the recording of the experiment. Subjects were instructed to watch a game video clip and recall player positions by writing them down on a blank on-screen basketball court schematic using simple signs: crosses and circles. They completed 6 sub-tasks for each low, medium and high level of mental demand, with a few minutes break between each level. All participants completed 8 sessions in different days, and here we consider one of the sessions (7) for data analysis[1].

Cognitive Load Modulation

Task difficulty levels were varied by the number of player positions to be recalled. In the low cognitive load level, 3 player positions were required, while 6 positions were required by the medium level and all 10 positions in the high level.

Apparatus

Eye activity was monitored using an ASL Eye-Trac 6 head mounted eye tracker system. Subjects were free to move their head but instructed to keep their eyes within the screen display range. Data was collected in a scene video and an eye video, where the scene and dynamic pupillary responses respectively were recorded at 15 frames per second.

Data Reduction and Variables

The data from six subjects for low and medium levels were used[1]. Measures of blink, pupillary response, fixation and saccade were processed from the eye video using scripts developed in MATLAB. Extracted data were superimposed on the eye video and played back to manually ensure that all features extracted correctly represented actual eye activities.

Blinks were identified as samples for which the pupil diameter was blocked by nearly half, until fully closed and reopened to above half-open. Pupil diameter during blinking was linearly interpolated between 2 frames before blink and 2 frames after blink. Under normal conditions, blink duration is around 100 - 150 ms and the 2 frames (133 ms) before and after each blink are long enough to estimate the pupil size occluded by blinking. Pupil size was measured in pixels and was filtered by a low pass filter to remove drift, tremors and other noise introduced in the measure. Fixation was defined as the eye position within 1 degree of the visual angle for at least 200 ms. Saccades and fixations were separated automatically using dispersion-based algorithms.

RESULTS

Subjective Rating

As expected, subjective ratings using a 9-point rating scale showed recalling 3 player positions (M=1.67, SD=0.52) is easier than recalling 6 player positions (M=3.33, SD=1.03). A paired two-tailed t-test conducted on these data showed a significant main effect (t(5)=3.95, p=0.01).

Eye Activity Measure

Eight dependent variables were employed to measure the mental effort: blink latency, blink rate, average pupil size in the time between 2s preceding and after the game video ended, standard deviation of pupil size in the 4-second period, fixation time, fixation rate, saccade size and saccade speed. Their trends are shown in Figure 1.

Overall-task Calibration and First-task Calibration

Due to the different feature value ranges for each participant, we experimented with overall-task calibration and first-task calibration method to minimize the between-subject variability in different mental effort scales. That is, we use the maximum and minimum parameter value ($V_{max,}V_{min}$) from all sub-tasks or estimate them from the first sub-task (increase/decrease by 50%) for each subject, then apply a normalization to the average feature value of 6 sub-tasks. The formula used is:

$$V_{cal} = \frac{V_{raw} - V_{\min}}{V_{\max} - V_{\min}}.$$

The effect of the overall-task calibration on the analysis is shown in the rightmost two columns of Table 1. The first-task calibration method gives similar levels of significance except for blink rate, which is significant only at 90% confidence.

[1] Video data for many other sessions, subjects and levels were unfortunately found to be corrupted, hence this subset.

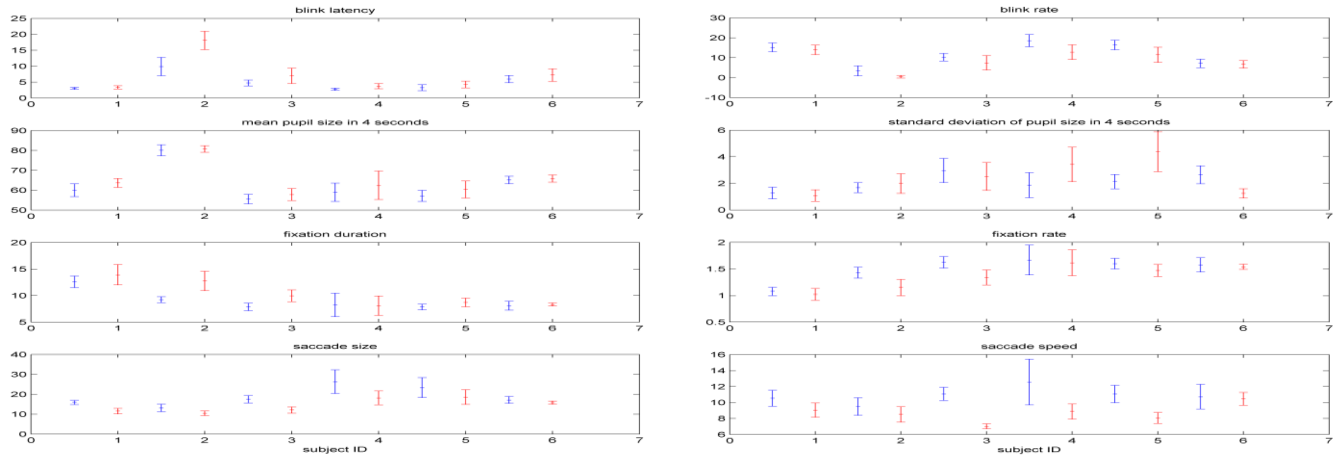


Figure 1. Raw feature value vs. subject ID for low (blue) and medium (red) mental effort levels. Bars display the feature value mean and range.

Table 1. Calibration. Paired t test analysis for the eight parameters, before and after (*) overall-task calibration.

Paired t-test	t(5)	P_{value}	$t(5)^*$	P_{value}^*
Blink latency	1.9450	0.1094	6.8591	0.0010
Blink rate	3.6409	0.0149	4.9380	0.0043
Mean Pupil size	4.2283	0.0083	5.2053	0.0035
Std pupil size	0.6282	0.5575	0.2185	0.8357
Fixation time	2.4114	0..0608	3.0688	0.0278
Fixation rate	2.9531	0.0318	3.4011	0.0192
Saccade size	4.6870	0.0054	6.7197	0.0011
Saccade speed	3.5623	0.0162	4.3702	0.0072

Correlation Coefficient between All Pairs of Features

As expected, parameters in same class of features are highly correlated except mean and standard deviation of pupil size. Blink, pupil size and eye movement features are uncorrelated with each other, as shown in Table 2.

DISCUSSION

In regards to blink activity, both blink latency and blink rate display clear mental effort related variations. Pupil size was measured from 2 seconds before and 2 seconds after the clip, involving mostly recall in this period, during which sustained working memory is heavily involved. The average pupil size for the two difficulty levels shows a significant effect as opposed to the standard deviation of pupil size, which indicates that in some cases pupil size is larger in a more difficult task level but shows less fluctuation. Meanwhile, fixation duration and fixation rate results indicate that significantly more attention was needed when the task was more complex. In addition, saccade speed and especially saccade size appear to have been highly discriminatory parameters.

These patterns of eye activity offers further insight into human mental effort. As more working memory and attentional resources are required in order to achieve high task performance, participants increased their blink latency, pupil size and fixation duration, and at the same time decreased their blink rate, fixation rate, saccade speed and saccade size. These patterns agree with previous literature, which suggest that blink and pupillary response can be an indicator of workload [1,2,5,6,7]. Interestingly, in this visual task study, eye movement appears to be a very suitable index of mental effort as well. However, these results contrast with those presented in two other studies [5,6] where fixation time, saccade size and saccade speed appeared to have no systematic changes with increasing workload in a visuospatial mock warfare task. The reason could be due to the nature of their task, depending on the degree of eye movement required to complete the task.

Table 2. Average correlation coefficients (Low in blue and medium in red; B1:blink latency; B2:blink rate; P1:mean pupil size; P2:standard deviation of pupil size; F1:fixation time; F2:fixation rate; S1:saccade size; S2:saccade speed).

L/M	B1	B2	P1	P2	F1	F2	S1	S2
B1		-0.86 -0.91	-0.20 -0.49	0.33 0.01	-0.02 -0.04	-0.02 0.12	-0.14 -0.39	0.10 -0.12
B2			0.34 0.38	-0.37 -0.02	-0.00 0.06	0.10 -0.13	0.22 0.48	-0.08 0.08
P1				-0.22 0.28	-0.19 -0.20	0.18 0.15	-0.00 0.44	0.07 0.47
P2					0.04 -0.28	-0.09 0.24	-0.04 0.14	0.17 0.12
F1						-0.96 -0.97	-0.12 -0.29	-0.32 -0.33
F2							0.15 0.23	0.33 0.36
S1								0.60 0.47

In this study, the higher load task of recalling 6 player positions requires an increased allocation of working memory resources, as well as identifying an increased number of defenders or attackers and making more selections and decisions while maintaining attention. This is highly demanding, as the basketball players are constantly moving in the game video. Some other tasks adopted previously for inducing different levels of mental effort, for instance, mental arithmetic and digit span tasks, depend on working memory heavily, while vigilance tasks need less working memory involvement [7]. It is hypothesized that a combination of eye blink, task-invoked pupillary response and eye movement each differently reflect the cognitive activities, since they are controlled by different nervous systems, i.e. the central or peripheral nervous system, and they may provide complementary information about mental effort. This is supported by the correlation analysis between the various pairs of features.

In this study, we proposed first-task calibration method, which compares well with calibration over the entire task. Although the aim of calibration is to reduce individual differences, it might also be used to find reference patterns that can distinguish the components of physical aspects of the task from the indicators of mental effort, e.g. remove the minimum value associated with basic eye function required from the feature values. Future work will evaluate the proposed methods on a much wider range of subjects.

IMPLICATIONS FOR INTERFACE DESIGN

Mental effort or cognitive load indices based on eye-activity form part of an area of active current interest in HCI. One proposed method employs eye movement to characterize system features during evaluation in order to assess interface quality [8]. There are also reports indicating that mental effort reflects user satisfaction and engagement [9]. One implication from this study is that eye activity can be employed as a potential tool to measure human mental effort in realistic human computer interaction scenarios. Multiple eye activity based features might be used more generally to provide improved load level discrimination. In HCI, intelligent interfaces will not only need to know where the attention of users is directed, but also how much of the user's working memory the interface/interaction/task is occupying. To evaluate an interface design, measuring the user's mental effort in real time via eye activity may be more objective, convenient and detailed, and less obtrusive method than self-report or dual task methodologies.

CONCLUSION

Different eye activity features spanning blink, task-invoked pupillary response and eye movement have been shown to each provide significant discriminative power between two levels of induced mental effort in a computer based training task typical of a semi-realistic training interface. Combination of these features has a distinct advantage as an objective measure of human mental effort, as different inhibitory mechanisms require mental effort for eye functions that, when combined, provide rich and possibly complementary information about mental effort. This is supported by the results of the correlation analysis between the various pairs of features. In turn, we may able to improve our understanding of human mental effort in real time, which may prove significant in the design and evaluation of usable, intelligent adaptive interfaces. Eye activity in HCI research has been mainly focused on eye movement as an input in human computer dialogue and for usability measurement, and both show promise but have yet to become widely used [10]. Although understanding of the effect of cognitive load on eye activity is growing, measurement systems have yet to be prototyped. Future work will focus on prototypes that benefit the interaction between users and computer systems.

ACKNOWLEDGMENTS

Our thanks to Damian Farrow and Tara Handke from Australian Institute of Sport and to all our participants. This research was supported by the Asian Office of Aerospace Research and Development, grant FA2386-10-1-4029.

REFERENCES

1. Irwin, D. E., Thomas, L. E., 2010. Eyeblinks and Cognition. *In:* Coltheart, V. (ed.) *Tutorials in Visual Cognition.* New York, London: Psychology Press, Taylor & Francis Group.
2. Kramer, A. F., 1991. Physiological metrics of mental workload: A review of recent progress. *In:* Damos, D. L. (ed.) *Multiple-task Performance.* London: Taylor & Francis Ltd.
3. Theeuwes, J., Belopolsky, A., 2010. Top-Down and Bottom-Up Control of Visual Selection Controversies and Debate. *In:* Coltheart, V. (ed.) *Tutorials in Visual Cognition.* New York, London: Psychology Press, Taylor & Francis Group.
4. Paas, F., Tuovinen, J. E., et al., 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist,* 38**,** 63 - 71.
5. Van Orden, K. F., Limbert, W., et al., 2001. Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 43**,** 111-121.
6. Greef, T., et al., 2009. Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation. *In: Proceedings of the 5th International Conference on Foundations of Augmented Cognition* (2009),219-228.
7. Klingner, J., Tversky, B., et al., 2010. Effects of visual and verbal presentation on cognitive load in vigilance, memory and arithmetic tasks. *Psychophysiology.*
8. Goldberg, J. H., Kotval, X. P.,1999. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6): 631-645.
9. Schmutz, P., Heinz, S., et al.,. 2009. Cognitive load in ecommerce applications: measurement and effects on user satisfaction. *Adv. in Hum.-Comp. Int.*(2009), 1-9.
10. Jacob, R. J. K., Karn, K. S., 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. *The Mind's eye: Cognitive and Applied Aspects of Eye Movement Research*, 573-603.

Appendix D

Pupillary Response Based Cognitive Workload Measurement under Luminance Changes

Jie Xu, Yang Wang, Fang Chen, Eric Choi

National ICT Australia[1]
University of New South Wales
jie.jackxu@gmail.com, {yang.wang, fang.chen, eric.choi}@nicta.com.au

Abstract. Pupillary response has been widely accepted as a physiological index of cognitive workload. It can be reliably measured with remote eye trackers in a non-intrusive way. However, pupillometric measurement might fail to assess cognitive workload due to the variation of luminance conditions. To overcome this problem, we study the characteristics of pupillary responses at different stages of cognitive process when performing arithmetic tasks, and propose a fine-grained approach for cognitive workload measurement. Experimental results show that cognitive workload could be effectively measured even under luminance changes.

Keywords: Cognitive workload, eye tracker, luminance, pupillary response.

1 Introduction

Cognitive workload measurement plays an important role in various application areas involving human-computer interface, such as air traffic control, in-car safety and gaming [2]. By quantifying the mental efforts of a person when performing tasks, cognitive workload measurement helps predict or enhance the performance of the operator and system. Physiological measures are one class of workload measurement techniques, which attempts to interpret the cognitive processes through their effect on the operator's body state [5]. In the past, physiological measures usually entailed invasive equipment. With the advance of sensing technologies in recent years, the measuring techniques have become less intrusive, especially those through remote sensing. As a physiological index, eye activity has been considered as an effective indicator of cognitive workload assessment, as it is sensitive to changes of mental efforts. Eye activity based physiological measures [1] [3] [4], such as fixation and saccade, eye blink, and pupillary response, can be detected unobtrusively through remote sensing.

The fact that changes of pupillary response occur during mental activity has long been known in neurophysiology, and it has been utilized to investigate cognitive workload. In an early work, Beatty investigates the pupillary response through experiments that involve tasks of short-term memory, language processing, reasoning

[1] NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council.

and perception [1]. Pupillary response is shown to serve as a reliable physiological measure of mental state in those tasks. Usually head-mounted eye trackers are used to measure pupillary response during the task. It is not until recently that remote eye tracking has become a popular approach for cognitive workload measurement [6] [9]. In comparison with the head-mounted eye trackers, remote eye tracker enables the non-intrusive measurement of cognitive workload without interfering with user's activity during the tasks. Moreover, remote video eye tracker is shown to be precise enough for measuring the pupillary response.

Though empirical evidence from the literature has demonstrated that eye activity based physiological measure is a useful indicator of mental efforts, it could be influenced by noise factors unrelated to the cognitive task. For example, it is reported that pupillary response is sensitive to illumination condition, fatigue, and emotional state [7] [8] [12]. These factors restrict the practical usage of pupillary response for cognitive workload measurement. In this paper we investigate the feasibility of measuring cognitive workload based on pupillary response even under luminance changes.

2 Related Work

As a non-intrusive means of measuring cognitive workload, remote eye tracker has been demonstrated to be precise enough for recording detailed information of pupillary response. Klingner *et al.* examine the pupil-measuring capability of video eye tracking in [6]. In their experiment, cognitive workload is measured using a remote video eye tracker during tasks of mental multiplication, short-term memory, and aural vigilance. It has been observed that the remote eye tracker can detect subtle changes in pupil size induced by cognitive workload variation. Similarly, Palinko *et al.* also use remote eye tracking to measure cognitive workload in their experiment [9]. In a simulated driving environment, pairs of subjects are involved in spoken dialogues and driving tasks. The driver's cognitive workload is estimated based on the pupillometric measurement acquired from the remote eye tracker. The pupillometric measurement and the driving performance exhibit significant correlation, which suggests the effectiveness of cognitive load measurement by remote eye tracker.

Although the physiological measure based on remote eye tracker has exhibited its usage for cognitive workload measurement, its performance could be affected by various noise factors. Luminance condition is especially known as an important factor that influences the pupil size. Pomplun and Sunkara compare the effects of cognitive workload and display brightness on pupil dilation and investigate the interaction of both factors in [10]. They design a gaze-controlled human computer interaction task that involves three levels of task difficulty. In the experiments, each level of the difficulty is combined with two levels of background brightness (black and white), which results in six different trial types. The experiment results show that the pupil size is significantly influenced by both the task difficulty and the background brightness. There is a significant increase of pupil size when the workload demand becomes higher under both background conditions. However, the pupil size corresponds to the highest workload under white background is even smaller than that corresponds to the lowest workload under black background.

3 Experiment

3.1 Participants and apparatus

Thirteen 24-to-46-year-old male subjects have been invited to participate in the experiment. All the subjects have normal or corrected-to-normal vision. Each subject receives a small-value reward for his participation.

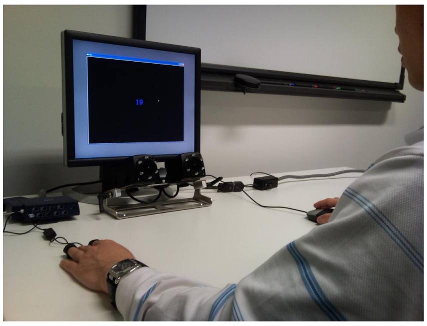


Figure. 1. Experiment setup.

The pupillary response data of each subject is recorded with a remote eye tracker (faceLAB 4.5 of Seeing Machines Ltd), which operates at a sampling rate of 50 Hz and continuously measures the subject' pupil diameters. The skin conductance data is also recorded with a galvanic skin response (GSR) sensor (ProComp Infiniti of Thought Technology Ltd). However the analysis of the GSR data is out of the scope of this paper. Visual stimuli are presented on a 21-inch Dell monitor with a screen resolution of 1024 by 768 pixels. The experiment setup is demonstrated in Figure 1.

3.2 Experiment design

Each subject is requested to perform arithmetic tasks under different luminance conditions. The arithmetic tasks have 4 levels of difficulty, and each level of task difficulty is combined with 4 levels of background brightness, which results in 16 different trial types in total.

For each arithmetic task, each subject is asked to sum up 4 different numbers sequentially displayed on the center of the screen, and then choose the correct answer on the screen through mouse input. The task difficulty depends on the range of numbers. For the first (lowest) difficulty level, each number is binary (0 or 1); for the second difficulty level, each number has 1 digit (1 to 9); for the third difficulty level, each number has 2 digits (10 to 99); for the fourth (highest) difficulty level, each number has 3 digits (100 to 999). Each number will be displayed for 3 seconds, and there is no time constraint for choosing the answer. Before the first number appears, different number of “X” will be displayed at the center of the screen for 3 seconds. The number of “X” corresponds to the number of digits for each arithmetic task.

During the experiment, the luminance condition varies when each subject performs arithmetic tasks. To produce different levels of luminance condition, luminance (grayscale value) of the background are set as 32, 96, 160 and 224 for the four levels of background brightness (L1, L2, L3, and L4), respectively. Black background will be displayed for 6 seconds before each arithmetic task. The time setting for each arithmetic task is depicted in Figure 2.

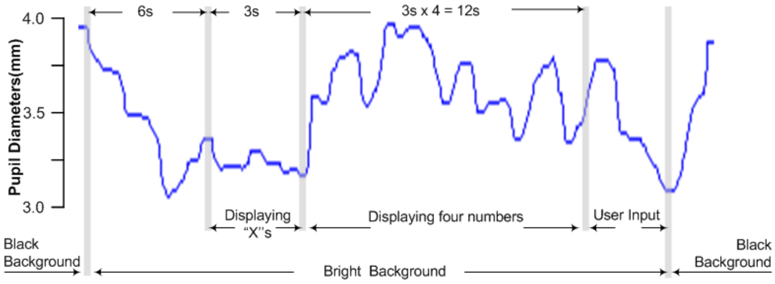


Figure 2. Time setting of an arithmetic task.

The experiment starts with a practice trial of which the data is not analyzed. Subsequently a one-minute resting data with black background is recorded before the test trials start. There are two tasks for each trial type, which results in 32 arithmetic tasks for each subject in the experiment. The tasks are presented randomly during the experiment. Once the subject finishes all the tasks, another one-minute resting data is also recorded. The whole experiment lasts about 25 minutes for each subject.

4 Analysis

In this section we analyze the correlation of the pupillary response and cognitive workload under different luminance conditions from the experimental data.

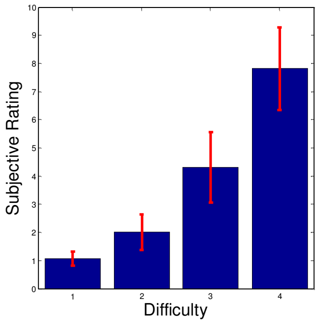


Figure 3. Subjective rating of task difficulty.

Figure 3 shows average subjective rating scores for the four levels of task difficulty from all the subjects. The scores range from 1 to 9, which correspond to the easiest and hardest tasks respectively. It can be known that there are significant differences between the subjective ratings of different task difficulty levels (F=108.63, $p<0.05$ in ANOVA test), which indicates the overall effectiveness of cognitive workload manipulation in the experiment.

4.1 Coarse-grained Analysis

For each subject, the pupillary response data of every arithmetic task during the experiment is examined. As a coarse-grained analysis, the average pupil diameter from the whole task period is used to characterize the cognitive workload. Figure 4 shows the average pupil diameters from that period under different levels of task difficulty and background brightness. It can be seen from the figure that the pupil diameter is influenced by the background brightness, in the sense that a smaller pupil diameter is usually observed under brighter background. On the other hand, the pupil diameter is also influenced by cognitive workload. For each background brightness level, the pupil diameter often increases when the task difficulty level becomes high. Together background brightness and cognitive workload could affect the pupil diameter. It can be observed that the pupil diameter at the highest task difficulty with highest background brightness is, in fact, smaller than that at the lowest task difficulty with lowest background brightness. This observation is consistent with previous empirical study that, luminance conditions take priority over cognitive demands in pupil diameter changes. Thus it is difficult to directly use the average pupil size or dilation to measure cognitive workload in the experiment.

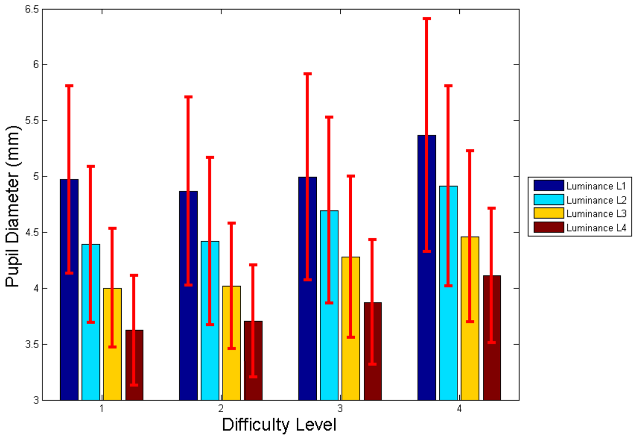


Figure 4. Pupil diameter under different task difficulty levels and background brightness conditions.

The above analysis shows that the coarse-grained measures of pupillary response could not effectively measure cognitive workload under luminance changes. To overcome this problem, we propose a fine-grained analysis of pupillary response in the following section. It is expected that the dynamic characteristics of cognitive process could be reflected by the fine-grained measures of pupillary response, which will improve cognitive workload measurement under complex environments.

4.2 Fine-grained Analysis

For a fine-grained analysis of pupillary response, the 12-second task period is divided into smaller-size intervals. As shown in Figure 5, we examine five 3-second intervals corresponding to different stages of the cognitive process when performing the task. We denote X as the interval for the "X" displaying interval, and N1, N2, N3 and N4 as the four 3-second number displaying intervals respectively. Additionally, we also examine two 6-second intervals based on N1, N2, N3 and N4. Let M1 be the first 6-second of number displaying interval, and M2 the second 6-second interval. The setting of task intervals can be found in Figure 5. On the basis of these interval definitions, there are 6 intervals for each arithmetic task. We measure the average pupil diameters from these 6 intervals.

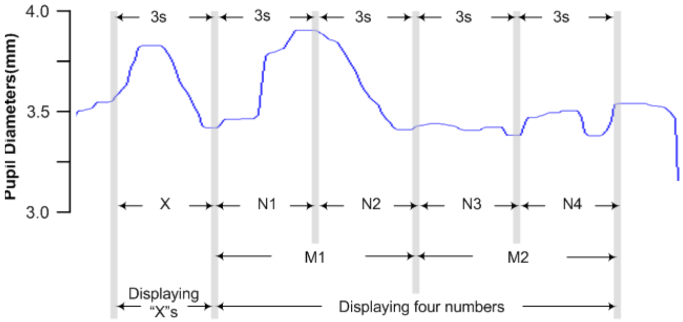


Figure 5. The setting of task intervals for fine-grained analysis.

For each task, average pupil diameter is measured for all the intervals. To reduce the influence of luminance condition, we normalize the measurement values for interval N1, N2, N3, N4, M1, and M2 using average pupil diameter d_X from the X interval, as there is no cognitive workload involved in that interval. Let d_N be one pupil size measurement from one task interval, its normalized measurement value is defined as $d_n = \frac{(d_N - d_X)}{d_X}$.

Figure 6 demonstrates the distributions of measurement values from M1 and M2 under different task difficulty levels. The figure demonstrates the characteristics of pupillary response at different time intervals. Even under the influence of luminance changes, the measurement values increase as the task difficulty increases. Such trend is more significant in the measurement values from M2 (F=3.93, p<0.05 in ANOVA test). We further examine interval M2 by studying the measurement values from N3 and N4, which are shown in Figure 7. As shown in Figure 7, the trend of the increase in the measurement values with increasing task difficulty is more significant in N4 (F=3.43, $p<0.05$).

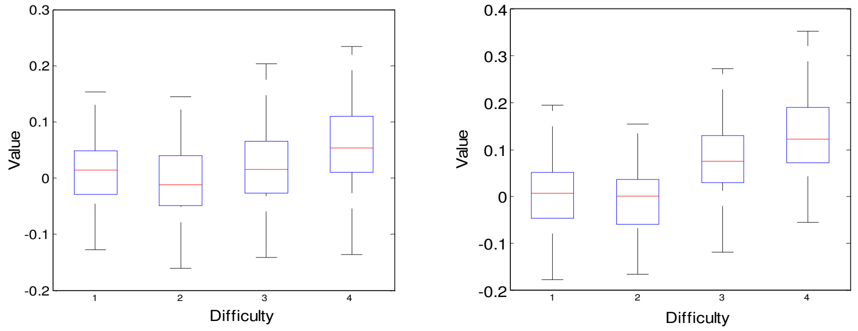


Figure 6. Box plots of measurement values (sample minimum, lower quartile, median, upper quartile, and maximum) under different task difficulty levels: (left) M1, (right) M2.

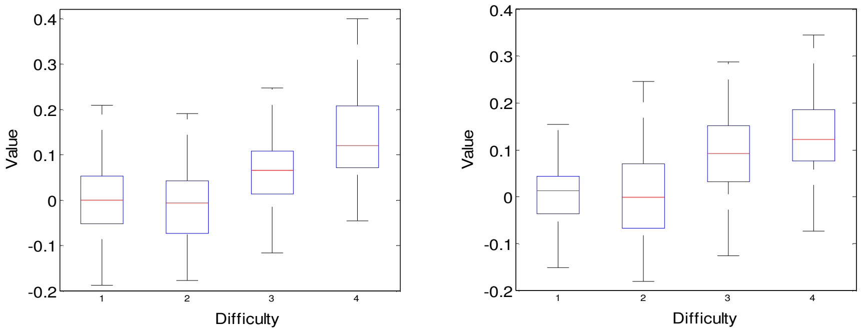


Figure 7. Box plots of measurement values (sample minimum, lower quartile, median, upper quartile, and maximum) under different task difficulty levels: (left) N3, (right) N4.

In addition to the above analysis, cognitive workload classification has also been investigated using measurements from different intervals. We employ a decision tree-based classification scheme [11] to classify the cognitive workload. Specifically, given the measurement values from different classes, a threshold is estimated such that maximum information gain can be achieved by splitting the data using that threshold. One threshold is needed for two-class classification while three thresholds are required for the four-class classification. We conduct both two-class classification (task difficulty 1, 2 vs. task difficulty 3, 4) and four-class classification of cognitive workload. The classification results are shown in Table 1. As shown in Table 1, M2 outperforms M1 for both two-class and four-class classification. N4 achieves the highest performance in both tasks. The measurements from different intervals reveal the dynamic characteristics of pupillary response at different stages of cognitive process, which can be utilized to improve the performance of cognitive workload assessment under complex environments.

Table 1. The classification results of different pupillary response measurements.

Pupillary Measurements	M1	M2	N3	N4
Two-class Classification	**59.3%**	**71.6%**	**68.9%**	**72.7%**
Four-class Classification	**36.6%**	**41.7%**	**43.0%**	**43.9%**

6 Conclusion

This work investigates the measurement of cognitive workload through remote eye tracking under the influence of luminance condition. We study the characteristics of pupillary response, by analyzing the measurements acquired from different stages of cognitive process. The experimental results demonstrate the feasibility of cognitive workload measurement under complex environments using the proposed fine-grained analysis. Our future work will be applying machine learning techniques to improve fine-grained analysis for cognitive workload measurement.

References

1. J. Beatty: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, vol. 91, pp. 276–292 (1982)
2. B. Cain: A review of the mental workload literature, Technical Report, Defence Research and Development Canada Toronto (2007)
3. S. Chen, J. Epps, N. Ruiz, F. Chen: Eye activity as a measure of human mental effort in HCI. International Conference on Intelligent User Interfaces, pp. 315–318 (2011)
4. C. Fogarty, J. Stern: Eye movements and blinks: Their relationship to higher cognitive processes. International Journal of Psychophysiology, vol. 8, pp. 35–42 (1989)
5. M. Grootjen, M. Neerincx, J. Weert: Task-based interpretation of operator state information for adaptive support. D. Schmorrow, M. Stanney, M. Reeves, (eds.) Foundations of Augmented Cognition (2nd edn.), pp. 236–242 (2006)
6. J. Klingner, R. Kumar, P. Hanrahan: Measuring the task-evoked pupillary response with a remote eye tracker. Eye Tracking Research and Applications Symposium, pp. 69–72 (2008)
7. A. Kramer: Physiological metrics of mental workload: A review of recent progress, D. Damos (ed.), Multiple-Task Performance, pp. 279–328, Taylor and Francis (1991)
8. S. Marshall: The index of cognitive activity: Measuring cognitive workload. IEEE Human Factors Meeting, pp. 7-5–7-9 (2002)
9. O. Palinko, A. Kun, A. Shyrokov, P. Heeman: Estimating cognitive load using remote eye tracking in a driving simulator, Eye Tracking Research and Applications Symposium, pp. 141–144 (2010)
10. M. Pomplun, S. Sunkara: Pupil dilation as an indicator of cognitive workload in human-computer interaction. International Conference on Human-Computer Interaction, pp. 542–546 (2003)
11. J. Quinlan: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993)
12. J. Xu, Y. Wang. F. Chen, H. Choi, G. Li, S. Chen, S. Hussain: Pupillary response based cognitive workload index under luminance and emotional changes. Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1627–1632 (2011).

Appendix E

Pupillary Response Based Cognitive Workload Index under Luminance and Emotional Changes

Jie Xu
National ICT Australia
Eveleigh, NSW 1430, Australia
jie.xu@nicta.com.au

Yang Wang
National ICT Australia
Eveleigh, NSW 1430, Australia
yang.wang@nicta.com.au

Fang Chen
National ICT Australia
Eveleigh, NSW 1430, Australia
fang.chen@nicta.com.au

Ho Choi
National ICT Australia
Eveleigh, NSW 1430, Australia
Eric.choi@nicta.com.au

Guanzhong Li
National ICT Australia
University of NSW
Eveleigh, NSW 1430, Australia
guanzhong.li@nicta.com.au

Siyuan Chen
National ICT Australia
University of NSW
Eveleigh, NSW 1430, Australia
siyuan.chen@nicta.com.au

Sazzad Hussain
National ICT Australia
University of Sydney
Eveleigh, NSW 1430, Australia
sazzad.hussain@nicta.com.au

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract

Pupillary response has been widely accepted as a physiological index of cognitive workload. It can be reliably measured with video-based eye trackers in a non-intrusive way. However, in practice commonly used measures such as pupil size or dilation might fail to evaluate cognitive workload due to various factors unrelated to workload, including luminance condition and emotional arousal. In this work, we investigate machine learning based feature extraction techniques that can both robustly index cognitive workload and adaptively handle changes of pupillary response caused by confounding factors unrelated to workload.

Keywords

Cognitive workload, eye tracker, feature extraction, machine learning, pupillary response

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User interfaces – Evaluation/methodology

General Terms

Human Factors, algorithms, experimentation

Introduction

Cognitive workload evaluation is an important issue in various research and application areas related to human-computer interaction such as adaptive automation and training, air traffic control, performance prediction, and driver safety [3]. With the advance of modern sensing technologies, more and more physiological measures have been developed for the assessment of cognitive workload. Among these techniques, video camera or imaging sensor based workload measures, especially those through remote sensing, have attracted increasing attention since they can provide physiological index of cognitive state in a non-intrusive way.

One valuable type of physiological measure involves workload effects on activities of human eye. The correlation between pupillary response and changes in mental workload has been studied for decades. In an early study [1], Beatty investigated task-evoked pupillary response in experiments containing various tasks such as language processing, reasoning, and perception. Pupil dilation was demonstrated to be a reliable physiological measure of mental state or processing load during the tasks. More recently, pupil measurement through video-based eye tracking has become a popular approach for cognitive workload evaluation due to its sensitivity and convenience [4,9]. The video information is acquired in a non-intrusive (particularly with remote systems) and continuous way without interfering with the user’s activity during the tasks. Moreover, the video sequences of eye tracking data can be captured with high frame rates (more than 30 frames per second) and processed in real-time.

Although empirical evidence from a number of studies has shown that eye-activity based physiological measures can be used as an effective indicator for increases in mental workload, the measures may fail to evaluate workload under complex environments, due to confounding (or noisy) factors unrelated to the cognitive task. Pupil dilation is known to be affected by both the illumination condition of the visual field and the emotional status of the subject [7,10]. For example, [5] reported the failure of workload measures due to changes in ambient illumination or screen luminance, which might give rise to greater variation of pupil size. Such factors restrict the usage of pupillary response as a workload index in practice. With machine learning based feature extraction techniques [11], in this work we investigate the feasibility of robustly measuring cognitive workload (with predefined workload levels) through remote eye tracking even under changes of luminance condition and emotional arousal.

Related Work

Recent cognitive studies have demonstrated the reliability of physiological measures obtained through remote eye tracking when the luminance condition is well-controlled. Klingner et al. examined the pupil measuring capability of video-based eye tracker for cognitive workload evaluation [4]. In the experiments, several arithmetic and memory tasks were performed by the subjects. Subtle changes in the task-evoked pupillary response were detected using a remote eye tracker. It was found that compared to previous obtrusive pupil measuring devices, the remote eye tracker could effectively measure the cognitive workload. In another experiment using dual task methodology, Palinko et al. also studied the pupillary response with a remote eye tracker [9]. The subjects performed both simulated vehicle driving and spoken dialogues. Pupil size data acquired from the remote eye tracker was used to evaluate the driver's cognitive workload. During the task, the physiological measure based on pupillary response exhibited significant correlation to those performance measures of driving.

Pupillary response is also influenced by luminance variations during experimental tasks. Pomplun and Sunkara investigated the effects of both cognitive workload and display brightness on pupil dilation in the experiment of a gaze-controlled human-computer interaction task [8]. In the visual task, three levels of task difficulty were combined with two levels of background brightness (black and white). The experimental results showed that under both black and white background conditions, the pupil area exhibited significant increase when workload demands became higher. However, under bright background even the pupil area corresponding to high level of task difficulty was significantly smaller than the pupil area corresponding to low level of difficulty under black background. In comparison with the task difficulty, the background brightness actually resulted in greater variation of the pupil area.

On the other hand, previous studies suggest that emotional arousal is another key factor affecting the pupillary response. Stanners et al. investigated pupillary response in tasks that involved both emotional and cognitive factors [10]. It was exhibited that cognitive demands took priority over emotional factors in modulating the pupillary response. With controlled luminance condition, Bradley et al. investigated the effects of emotional arousal on pupillary response during picture viewing [2], using a set of pictures selected from the International Affective Picture System (IAPS) [6]. It was found that pupillary changes became larger when viewing pleasant and unpleasant images.

Experiment

So far twelve 24-to-35-year-old male participants (20-30 participants in total are expected) have been invited to perform arithmetic tasks under changes of luminance condition and emotional arousal simultaneously. The whole experiment consists of three parts and lasts about 15 minutes. In the first part, the subject is asked to sum up numbers with blank background (black screen). In the second and third parts, the subject is asked to sum up numbers with pleasant and unpleasant background images shown on the screen. Different task difficulty levels and background conditions are employed to manipulate the cognitive workload, as well as background luminance and emotional arousal during the experiment.

For each arithmetic task, the subject is asked to sum up 4 different numbers sequentially shown at the center of the screen, and then choose the correct answer on the screen through mouse input. There are 4 levels of task difficulty depending on the range of numbers. For the first (lowest) difficulty level, each number is binary (0 or 1); for the second difficulty level, each number is 1-digit (1 to 9); for the third difficulty level, each number is 2-digit (10 to 99); for the fourth (highest) difficulty level, each number is 3-digit (100 to 999). Each number will be displayed for 3 seconds, and there is no time constraint for choosing the answer. Before the first number appears, an "X" will be displayed at the center of the screen for 3 seconds.

To vary both the luminance condition and emotional arousal, pleasant and unpleasant background images are shown on the screen when the subject performs arithmetic tasks in the second and third parts of the experiment. A background image will be displayed for 6 seconds before each arithmetic task. Subsequently, the subject will perform the arithmetic task with the background image remaining on the screen. Eight pleasant images (mean valence/arousal = 7.1, 5.7) and eight unpleasant images (mean valence/arousal = 2.8, 4.8) are selected from the IAPS database. The mean luminance (Y value) of the images ranges from 53 to 174.

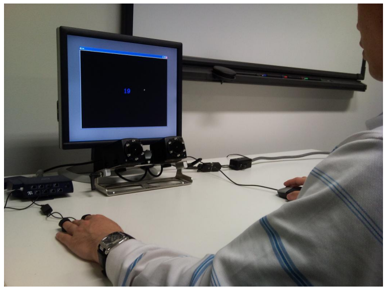


Figure 1. Experiment setup

One minute resting data with black screen is recorded at the beginning and the end of the whole experiment for each subject. There are 8 arithmetic tasks randomly given in each experiment part (2 for each difficulty level). During the experiment, pupillary response of each subject is recorded with a remote eye tracker (faceLAB 4.5 of Seeing Machines Ltd). Skin conductance is also recorded with a GSR sensor (ProComp Infiniti of Thought Technology Ltd).

Preliminary Analysis

Figure 2 shows the average pupil diameter under different task difficulty levels and background conditions (ignoring pupillary response during eye blinks). It can be seen that for the arithmetic tasks with the black background, the pupil diameter increases when the task difficulty level becomes high (F>11, p<0.01 in ANOVA test for pupil diameter). However, such relationship can no longer be observed for the tasks with background images. Both the luminance change of the screen and the emotional arousal when viewing the pictures appear to influence the pupillary response. Specifically, the pupil size recorded at the highest task difficulty level with background images is, in fact, smaller than the pupil size recorded at the lowest task difficulty level with black background. The phenomenon is consistent with previous empirical research showing that luminance conditions take priority over cognitive demands and emotional arousal in pupil size changes. Considering the influence of noisy factors unrelated to cognitive workload, pupil size or dilation may not effectively index cognitive workload under complicated situations.

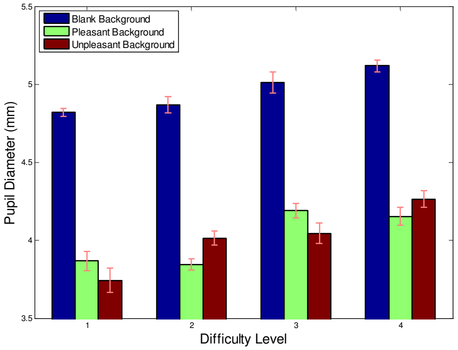


Figure 2. Pupil diameter under different task difficulty levels and background conditions

In this work, we investigate the feasibility of robustly measuring cognitive workload even under the effects of noisy factors including luminance changes and emotional arousal. The problem could be solved if the physiological features (measures) extracted from the pupillary response data could both characterize cognitive workload and adapt to changes caused by the confounding factors. To test its feasibility, a simple difference feature (the difference of the average pupil diameter between the first half and second half interval of the task) is employed to characterize the cognitive workload changes. The distribution of normalized feature values for different difficulty levels from all the pupillary response data is depicted in Figure 3. Even with changes of background luminance and emotional arousal, the feature value increases as the task difficulty level increases (F>8, p<0.01 in ANOVA test for the feature).

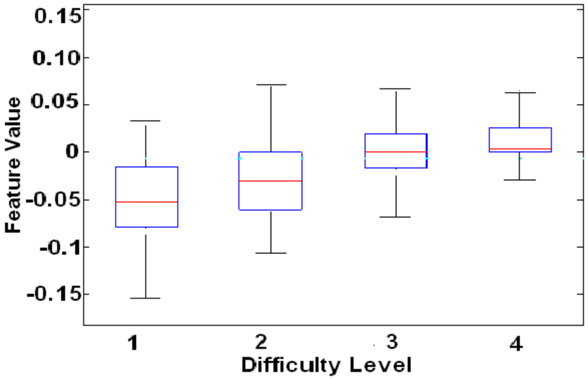


Figure 3. Box plot of feature values (sample minimum, lower quartile, median, upper quartile, and maximum) corresponding to different task difficulty levels

Using a decision tree-based classification scheme, about 49% overall accuracy of 4-class difficulty level estimation is achieved based on the difference feature. Meanwhile, with the same classification scheme only 28% accuracy is achieved by using the pupil dilation measure. The improvement shows the potential of robust workload measurement under the influence of confounding factors using pupillary response based feature. It should be noted that such feature is heuristically derived from manual inspection of the data itself, and the performance of workload evaluation could be boosted with machine learning based feature extraction techniques.

Ongoing Work

Our current work focuses on developing machine learning algorithms that can automatically find optimal features for robust workload measurement under noisy factors. There are quite a few systematic ways for solving this optimization problem, and Boosting is one popular algorithm that is suitable in this instance [11]. Boosting is a type of learning algorithm, which creates a classifier that can predict the labels of unseen data based on the given examples and their labels. In its original form, the Boosting algorithm is used to form a strong classifier from a set of weak classifiers. A strong classifier is defined as a classifier that correlates arbitrarily well with the true classification, whereas a weak classifier only correlates slightly with the true classification. However it can also be used as a feature selection scheme if we relate each weak classifier to a single feature. For example, we can define a weak classifier $h_j(x)$ that consists of a feature f_j, a threshold θ_j, and the parity $p_j = \pm 1$, which indicates the following simple classification rule: if $p_j f_j(x) \leq p_j \theta_j$ then $h_j(x) = 1$, otherwise $h_j(x) = 0$. The Boosting algorithm then creates a strong classifier $H(x) = \sum_j \alpha_j h_j(x)$ by selecting $h_j(x)$ iteratively from a pool of weak classifiers, and each $h_j(x)$ is weighted by α_j, which relates $h_j(x)$'s classification accuracy on the examples. Additionally, the examples are reweighted so that future weak classifiers can focus on the examples misclassified by the previous classifiers. In this work, each extracted feature can be viewed as a weak classifier consisting of a time interval vector $\overrightarrow{T_j}$, a threshold and also a parity value. Currently $\overrightarrow{T_j}$ is set heuristically using the first and second half of each task. To improve accuracy of cognitive workload indexing, the optimal set of weak classifiers (features) can be obtained through the Boosting algorithm. Furthermore, the recorded GSR data will also be analyzed in a similar way to complement the pupillary response analysis.

Summary

This work studies cognitive workload evaluation through remote eye tracking under the influence of confounding factors such as luminance condition and emotional arousal. We are employing Boosting based feature extraction to both robustly measure workload and adaptively handle changes of pupillary response caused by confounding factors. The proposed technique can be used in various applications involving cognitive workload evaluation under complex environments.

Acknowledgements

We would like thank Natalie Ruiz and Julien Epps for discussions of experiment design and data analysis. We wish to thank our pilot participants as well as future participants for their time and cooperation.

References

[1] J. Beatty: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, vol. 91, pp. 276–292, 1982.

[2] M. Bradley, L. Miccoli, M. Escrig, P. Lang: The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology, vol. 45, pp. 602-607, 2008.

[3] B. Cain, A review of the mental workload literature, Technical Report, Defence Research and Development Canada Toronto, 2007.

[4] J. Klingner, R. Kumar, P. Hanrahan: Measuring the task-evoked pupillary response with a remote eye tracker. Eye Tracking Research and Applications Symposium, pp. 69–72, 2008.

[5] A. Kramer: Physiological metrics of mental workload: A review of recent progress, D. Damos (ed.), Multiple-Task Performance, pp. 279–328, Taylor and Francis, 1991.

[6] P. Lang, M. Bradley, B. Cuthbert: International affective picture system: Affective ratings of pictures and instruction manual. Technical Report, University of Florida, 2005.

[7] S. Marshall: The index of cognitive activity: Measuring cognitive workload. IEEE Human Factors Meeting, pp. 7-5–7-9, 2002.

[8] M. Pomplun, S. Sunkara: Pupil dilation as an indicator of cognitive workload in human-computer interaction. International Conference on Human-Computer Interaction, pp. 542–546, 2003.

[9] O. Palinko, A. Kun, A. Shyrokov, P. Heeman: Estimating cognitive load using remote eye tracking in a driving simulator, Eye Tracking Research and Applications Symposium, pp. 141–144, 2010.

[10] R. Stanners, M. Coulter, A. Sweet, P. Murphy: The pupillary response as an indicator of arousal and cognition, Motivation and Emotion, vol. 3, pp. 319-340, 1979.

[11] R. Schapire: The Boosting approach to machine learning：An overview, D. Denison, M. Hansen, C. Holmes, B. Mallick, B. Yu (ed.), Nonlinear Estimation and Classification, Springer, 2001.